NAVAL RESEARCH LOGISTICS QUARTERLY

DECEMBER 1974
VOL. 21, NO. 4

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $10.00 a year in the U.S. and Canada, $12.50 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, NAVEXOS P-35

ON THE OPTIMALITY OF THE BAYES PREDICTION POLICY IN TWO-ECHELON MULTISTATION INVENTORY MODELS*

S. Zacks

Case Western Reserve University

ABSTRACT

The present paper provides a proof that the Bayes prediction ordering policy developed in [3] is an optimal policy, in the sense that it minimizes the total expected discounted cost of ordering for any finite planning horizon.

1. INTRODUCTION

In previous studies on the two-echelon multistation inventory model [2, 3] we have developed an ordering policy for the upper echelon, which is called a Bayes prediction policy. According to this policy the upper echelon orders at the beginning of each month a quantity Z_n, which is the smallest nonnegative integer that satisfies a certain probability constraint. We have assumed a 2-month lead time of the upper echelon orders, and thus the probability constraint is such that the anticipated stock in the upper echelon, at the beginning of the $(n+2)$nd month is not smaller than a preassigned confidence level γ. In the present paper we prove that this Bayes prediction ordering policy is optimal in the sense that it minimizes the total expected discounted cost of ordering over any finite planning horizon.

Iglehart and Jaquette [1] have studied a similar problem, bur their model is substantially different from ours. In their model, if V_n is the quantity in stock at the beginning of the nth month then the demand X_n during the nth month has a distribution whose range is $[0, V_n]$ and it belongs to a family of total positivity two TP_2 (for example, the distribution of X_n given V_n may be a binomial with parameters θ, $0 < \theta < 1$, and V_n; where V_n is like the number of independent Bernoulli trails and θ is the probability of success). They have established the optimality of the policy which orders the smallest quantity for which the shortage probability does not exceed a given γ, $0 < \gamma < 1$. In our problem the demand from the upper echelon consists of the total orders from the lower echelon stations. The total of all the orders cannot exceed the stock level V_n, but on the other hand it is determined by a stock allocation algorithm (see [3]) for the lower echelon stations, and therefore follows particular probability distributions. For this reason we have to analyze the optimality question more specifically, and the result of Iglehart and Jaquette [1] cannot be applied directly. The main theoretical approach is, however, similar. Our framework is a Bayes adaptive one (see [3] for details) since we do not have complete information on the

*Research performed at the Program in Logistics of the Institute for Management Science and Engineering, The George Washington University under contract N00014-67-A-0214 with the Office of Naval Research.

demand distributions. In the following section we introduce the model and the basic tools. These are presented without details. The details can be found in the previous studies [2, 3]. The proof of the main result is given in sec. 3.

2. BASIC RELATIONSHIPS

In the model under consideration we have a depot (upper echelon) and s stations (lower echelon). The stock level at the depot at the beginning of the nth month is denoted by V_n. The stock levels at the stations, at the beginning of the nth month are $q_{n,1}, \ldots, Q_{n,s}$. The lower echelon stations compute at the beginning of the nth month their desirable (base) stock levels $k_{n,i}\,(i=1, \ldots, s)$ for the beginning of the $(n+1)$st month. These parameters depend on certain cost ratios and on the anticipated distributions of demand. According to the values $k_{n,1}, \ldots, k_n, s$, and V_n ordering levels $Y_{n,i}$ $(i=1, \ldots, s)$ are determined following the allocation algorithm which is presented in [3]. Let $a^+ = \max(0, a)$. The $Y_{n,i}$ values can be negative (shipment of stock back to the depot), and satisfy the conditions

$$Y_{n,i} \geq -Q_{n,i}, \qquad \text{for every } i=1, \ldots, s \tag{2.1}$$

$$\sum_{i=1}^{s} Y_{n,i}^{+} \leq V_n.$$

The qualities ordered by the lower echelon stations arrive at the stations from the depot within the same month period. They are available at the end of that same period. The upper echelon depot orders at the beginning of the nth month a quantity Z_n, which arrives at the depot at the end of the $(n+1)$st month. For more details see [3]. The demand distributions at the lower echelon stations are assumed to be Poisson with unknown means, θ_i. We have developed Bayes prediction policies which assigns these unknown means prior gamma distributions. The set of sufficient statistics for the sequential model are the vectors $\{\underline{T}_n;\ n=1, 2, \ldots\}$ where $T_{n,i} = \sum_{j=0}^{n} X_{j,i}$ is the total demand at the ith station during the first n-month. The sequence of n vectors $(\underline{T}_1, \underline{T}_2, \ldots, \underline{T}_n)$ is denoted by $\mathcal{T}_n$ $(n=1, 2, \ldots)$. We let $\mathcal{F}_n$ denote the σ-field generated by $(\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n)$. All the conditional probabilities which are discussed here are determined by the mixtures of the probabilities induced by the Poisson distributions, with respect to the posterior distributions of $(\theta_1, \ldots \theta_s)$ given $\mathcal{F}_n$. Let $W_n = \sum_{i=1}^{s} Y_{n,i}$ be the total demand from the depot, at the beginning of the nth month. According to the allocation algorithm, and since the quantities $k_{n,i}(i=1, \ldots, s)$ and $Q_{n,i}$ are $\mathcal{F}_{n-1}$ measurable, the conditional distribution of W_n given $\mathcal{F}_{n-1}$ is a one-point distribution whose formula is

$$P[W_n = w | \mathcal{F}_{n-1}] = I\left\{ w = \sum_{i=1}^{s} (k_{n,i} - Q_{n,i}),\ \sum_{i=1}^{s} (k_{n,i} - Q_{n,i})^+ \leq V_n \right\} \tag{2.2}$$

$$+ I\left\{ w = V_n,\ \sum_{i=1}^{s} (k_{n,i} - Q_{n,i})^+ > V_n \right\};$$

where $I\{\cdot\}$ is the indicator function of the set specified in the bracket. Moreover, $\{A, B\} = \{A \cap B\}$

is the intersection. Let B_{n-1} be a set of $\mathcal{T}_{n-1}$ sequences in $\mathcal{F}_{n-1}$ for which

$$\sum_{i=1}^{s} (k_{n,i} - Q_{n,i})^+ \leq V_n;$$

and let $\bar{B}_{n-1}$ denote the complement of B_{n-1}. The recursive relationship between V_{n+1} and V_n is given by:

$$V_{n+1} = \left[V_n + Z_{n-1} + \sum_{i=1}^{s} Q_{n,i} - \sum_{i=1}^{s} k_{n,i} \right]^+ I\{B_{n-1}\} + Z_{n-1} I\{\bar{B}_{n-1}\}, \tag{2.3}$$

where Z_{n-1} denotes the order issued by the depot at the beginning of the $(n-1)$st month. We have shown in [3] that a Bayes prediction ordering policy yields the quantities

$$Z_n = (\chi_{n,s}(\gamma) - V_{n+1})^+, \tag{2.4}$$

where $\chi_{n,s}(\gamma)$ is the γ-fractile of the total anticipated demand distribution of the s lower echelon stations during the $(n+1)$st and $(n+2)$nd month (see [3]). $\chi_{n,s}(\gamma)$ is an $\mathcal{F}_{n-1}$ measurable random variable. An ordering quantity $\tilde{Z}_n$ is called *admissible* if $\tilde{Z}_n \geq Z_n$, where Z_n is given by (2.4). Indeed, we impose the constraint that the shortage anticipated probability will not exceed $\alpha = 1 - \gamma$. For matter of convenience define

$$\psi_n = \chi_{n,s}(\gamma) - V_n - \sum_{i=1}^{s} Q_{n,i} + \sum_{i=1}^{s} k_{n,i}. \tag{2.5}$$

From (2.3)–(2.5) we obtain the recursive relation, with Z_0 arbitrary,

$$Z_n = (\psi_n - Z_{n-1})^+ I\{B_{n-1}\} + (\chi_{n,s}(\gamma) - Z_{n-1})^+ I\{\bar{B}_{n-1}\}, \qquad n = 1, 2, \ldots. \tag{2.6}$$

3. THE OPTIMALITY OF THE BAYES PREDICTION POLICY

Consider a planning horizon of $N+2$ months, where N is an arbitrary finite integer. Let β, $0 < \beta \leq 1$, be a discount factor. Our objective is to determine admissible $\tilde{Z}_n$ values so as to minimize the total expected discounted cost for the N orders. Suppose that the ordering cost is 1 cost unit per item ordered. Thus, the last order induces a cost of $\tilde{Z}_n$. Introduce the function

$$C_N^{(0)}(z) = z, \qquad z \geq Z_N \tag{3.1}$$

for any $N = 1, 2, \ldots$. Furthermore, according to (2.6) let

$$Z_n(z) = (\psi_n - z)^+ I\{B_{n-1}\} + (\chi_{n,s}(\gamma) - z)^+ I\{\bar{B}_{n-1}\}, \qquad n = 1, 2, \ldots. \tag{3.2}$$

Let $M_{N-n}^{(0)}(u)$ denote the minimal anticipated total discounted cost of the last $(n+1)$ orders, at the beginning of the $(N-n)$th month, where u is the quantity ordered at the beginning of the $(N-n-1)$st month. Then, by the principle of Dynamic Programming

$$M^{(n)}_{N-n}(u) = \min_{z \geq Z^{(u)}_{N-n}} \{z + \beta E_{N-n-1}\{M^{(n-1)}_{N-n+1}(z)\}\}, \tag{3.3}$$

where $E_{N-n-1}\{\cdot\}$ is the conditional expectation, according to the anticipated distribution, given $\mathcal{F}_{N-n-1}$. We notice from (3.1) that

$$M^{(0)}_{N}(u) = Z_N(u). \tag{3.4}$$

At the beginning of the $(N-1)$th month we have

$$M^{(1)}_{N-1}(u) = \min_{z \geq Z^{(u)}_{N-1}} \{z + \beta E_{N-2}\{Z_N(z)\}\}. \tag{3.5}$$

Let

$$C^{(1)}_{N-1}(z) = z + \beta E_{N-2}\{Z_N(z)\}. \tag{3.6}$$

According to 3.2, we have

$$C^{(1)}_{N-1}(z) = z + \beta E_{N-2}\{I\{B_{N-1}, \psi_N > z\}(\psi_N - z) + I\{\bar{B}_{N-1}, \chi_{N,s}(\gamma) > z\}(\chi_{N,s}(\gamma) - z)\}. \tag{3.7}$$

LEMMA: *Let*

$$C^{(1)}_{n}(z) = z + \beta E_{n-1}\{I\{B_n, \psi_{n+1} > z\}(\psi_{n+1} - z) + I\{\bar{B}_n, \chi_{n+1,s}(\gamma) > z\}(\chi_{n,s}(\gamma) - z)\}, \quad n = 1, 2, \ldots. \tag{3.8}$$

Then $C^{(1)}_n(z)$ *is a nondecreasing function of* z, *and is strictly increasing if* $\beta < 1$.

PROOF: Let $\underset{z}{\Delta} C^{(1)}_n(z) = C^{(1)}_n(z+1) - C^{(1)}_n(z)$. From (3.8) we obtain that

$$\underset{z}{\Delta} C^{(1)}_n(z) = 1 + \beta E_{n-1}\{I\{B_n, \psi_{n+1} > z+1\} \cdot (\psi_{n+1} - z - 1) - I\{B_n, \psi_{n+1} > z\}(\psi_n - z) + I\{\bar{B}_n, \chi_{n+1,s}(\gamma) > z+1\}(\chi_{n+1,s}(\gamma) - z - 1) - I\{\bar{B}_n, \chi_{n+1,s}(\gamma) > z\}(\chi_{n+1,s}(\gamma) - z)\}. \tag{3.9}$$

We notice that

$$I\{B_n, \psi_{n+1} > z+1\} - I\{B_n, \psi_{n+1} > z\} = -I\{B_n, \psi_{n+1} = z+1\}. \tag{3.10}$$

Similarly with the other terms. Hence,

$$\underset{z}{\Delta} C^{(1)}_n(z) = 1 - \beta E_{n-1}\{I\{B_n, \psi_{n+1} \geq z+1\} + I\{\bar{B}_n, \chi_{n+1,s}(\gamma) \geq z+1\}\}. \tag{3.11}$$

Moreover,

$$I\{B_n, \psi_{n+1} \geq z+1\} + I\{\bar{B}_n, \chi_{n+1,s}(\gamma) \geq z+1\} \leq I\{B_n\} + I\{\bar{B}_n\} = 1. \tag{3.12}$$

Therefore,

$$\underset{z}{\Delta} C_n^{(1)}(z) \geq 1-\beta. \tag{3.13}$$

This proves the lemma. (Q.E.D.)

Since $C_{n-1}^{(1)}(z)$ is nondecreasing, its minimum is attained at $Z = Z_{n-1}(u)$, and we have

$$M_{N-1}^{(1)}(u) = Z_{N-1}(u) + \beta E_{N-2}\{Z_n(Z_{N-1}(u))\}. \tag{3.14}$$

For the beginning of the $(N-2)$nd month,

$$M_{N-2}^{(2)}(u) = \min_{z \geq Z_{N-2}(u)} \{z + \beta E_{N-3}\{M_{N-1}^{(1)}(z)\}\}. \tag{3.15}$$

According to (3.6) and (3.14) we define

$$\begin{aligned} C_{N-2}^{(2)}(z) &= z + \beta E_{N-3}\{C_{N-1}^{(1)}(Z_{N-1}(z))\} \\ &= z + \beta E_{N-3}\{Z_{N-1}(z)\} + \beta^2 E_{N-3}\{Z_{N-1}(z)\} \\ &= C_{N-2}^{(1)}(z) + \beta^2 E_{N-3}\{Z_N(Z_{N-1}(z))\}. \end{aligned} \tag{3.16}$$

According to the Lemma, $C_{N-2}^{(1)}(z)$ is a nondecreasing function of z. Furthermore, according to (3.2), $Z_n(z)$ is a nonincreasing function of z for every $n = 1, 2, \ldots$ Hence, $Z_N(Z_{N-1}(z))$ is a nondecreasing function of z, and so is its expectation. Thus we have proven that $C_{N-2}^{(2)}(z)$ is a nondecreasing function of z, and the optimal value for the $(N-2)$nd month is $Z_{N-2}(u)$. In this manner, we proceed to prove by induction that the Bayes prediction policy, given by Z_n, is optimal for every $n = 1, \ldots, N$.

THEOREM: *For each* j = 1, . . . , N *define recursively*

$$C_{N-j}^{(j)}(z) = z + \beta E_{N-j-1}\{M_{N-j+1}^{(j-1)}(z)\}. \tag{3.17}$$

Then, $C_{N-j}^{(j)}(z)$ *is a nondecreasing function of* z *for every* j = 1, . . ., N, *and*

$$M_{N-j}^{(j)}(u) = C_{N-j}^{(j)}(Z_{N-j}(u)), \qquad j = 1, \ldots, N. \tag{3.18}$$

PROOF: We have established these properties of $C^{(j)}_{N-j}(z)$ and $M^{(j)}_{N-j}(u)$ for $j=1, 2$. Assume that it holds for all $j=1, \ldots, n$, $n=1, \ldots, N-1$. We prove it for $j=n+1$. According to (3.17) and the induction hypothesis

$$C^{(n+1)}_{N-n-1}(z) = z+\beta E_{N-n-2}\{C^{(n)}_{N-n}(Z_{N-n}(z))\}\}. \tag{3.19}$$

Applying (3.17) and (3.18) with respect to $C^{(n)}_{N-n}(Z_{N-n}(z))$., we can write

$$C^{(n+1)}_{N-n-1}(z) = z+\beta E_{N-n-2}\{Z_{N-n}(z)+\beta E_{N-n-1}\{C^{(n-1)}_{N-n+1}(Z_{N-n+1}(Z_{N-n}(z)))\}\}. \tag{3.20}$$

Or, according to the law of the iterated expectation and the definition of $C^{(1)}_{N-n-1}(z)$ we have

$$C^{(n+1)}_{N-n-1}(z) = C^{(1)}_{N-n-1}(z)+\beta^2 E_{N-n-1}\{C^{(n-1)}_{N-n+1}(Z_{N-n+1}(Z_{N-n}(z)))\}. \tag{3.21}$$

According to the Lemma, $C^{(1)}_{N-n-1}(z)$ is a nondecreasing function of z. According to the induction hypothesis $C^{(n-1)}_{N-n+1}(z)$ is a nondecreasing function of z. Moreover, $Z_{N-n+1}(Z_{N-n}(z))$ is a nondecreasing function of z, and so is its expectation. Thus, we have proven that $C^{(n+1)}_{N-n-1}(z)$ is a nondecreasing function of z. Finally, $C^{(n-1)}_{N-n-1}(z)$ is minimized by $Z_{N-n-1}(u)$, and (3.18) follows from this and (3.3). (Q.E.D.)

REFERENCES

[1] Iglehart, D. L. and S. C. Jaquette, "Optimal Policies under the Shortage Probability Criterion for an Inventory Model with Unknown Dependent Demand," Nav. Res. Log. Quart. *16*, 485–493 (1969).

[2] Zacks, S. and J. Fennel, "Bayes Adaptive Control of Two-Echelon Inventory Systems, I: Development for a Special Case of a One-Station Lower Echelon and Monte Carlo Simulation," Nav. Res. Log. Quart. *19*, 15–28 (1972).

[3] Zacks, S. and J. Fennel, (1973). "Bayes Adaptive Control of Two-Echelon Inventory Systems, II: The Multi-Station Case. Nav. Res. Log. Quart. *21*, 586–593 (1974).

BAYES ADAPTIVE CONTROL OF TWO-ECHELON INVENTORY SYSTEMS, II: THE MULTISTATION CASE*

S. Zacks
Case Western Reserve University

and

J. Fennell
The George Washington University

ABSTRACT

The present paper extends the results of [7] to cases of multistation lower echelon. For this purpose an algorithm for the optimal allocation of the upper echelon stock among the lower echelon stations is developed. The policy of ordering for the upper echelon is an extension of the Bayes prediction policy developed in [7]. Explicit formulae are presented for the execution of this policy. Several simulation runs are presented and analyzed for the purpose of obtaining information on the behavior of the system, under the above control policy, over short and long periods.

0. INTRODUCTION

The present paper constitutes the second part of our study on the two-echelon inventory problem. In Part I, which was recently published [7], we treated cases in which the lower echelon consisted of one station only. The present paper extends the previous results to the case of two or more stations in the lower echelon. The inventory model follows the same lines as in the previous studies [3] and [7], and the general objectives are the same. The monthly cost of the inventory system is the sum of the monthly costs at the individual stations due to overstocking or shortages. The leadtime of inventory replenishment is (up to) 1 month between echelons, and between 1 and 2 months from the outside resources (manufacturers) to the upper echelon (depot). According to this structure, the ordering of the lower echelon stations at the beginning of each month is designed so that the anticipated total cost of the inventory system during the following month is minimized. To attain this minimization, under the physical constraints of the stock available in the system, an algorithm for an optimal stock allocation is developed. The sequential ordering policy is myopic. It has been previously shown [3] that for the present inventory model the myopic procedure for the lower echelon is optimal. However, the minimization of the total anticipated cost depends on the quantities of stock available at the upper echelon at the beginning of each month. According to the cost structure of the present study this minimization requires large (infinite) quantities at the upper echelon. In order to make the problem meaningful we adopted the Bayes prediction principle (see also [7]). According to this principle the upper echelon

*Research performed at the Program in Logistics of the Institute for Management Science and Engineering, The George Washington University, under Contract N00014-67-A-0214, Task 0001, Project NR 347-020, Office of Naval Research.

orders at the beginning of each month the smallest quantity that can insure with (posterior) confidence probability γ $(0 < \gamma < 1)$ that all the orders of the lower echelon stations will be supplied. Such an ordering policy is called a *Bayes prediction policy*, since we employ Bayes adaptive techniques to estimate the unknown parameters. Explicit formulae for the determination of the ordering levels according to this policy have been developed and presented. We remark in this connection that the Bayes prediction policy is similar in essence to the "shortage probability" criterion, which was studied by Iglehart and Jaquette [2]. They have considered an inventory model in which the demand depends on the stock level at the system, and the system issues at the beginning of each month an order which is equal to the minimal amount so that the probability of shortage will not exceed some value, say α. One month leadtime is assumed in that study. It was established in [2] that if the distribution of the demand, given the stock level, belongs to a TP_2 (total positivity two) family then the above "shortage probability" policy is optimal. It minimizes the total expected discounted cost of *ordering* over a given planning horizon of N months. In our study the situation is much more complicated. We have been successful, however, in proving that the Bayes prediction policy is optimal in the sense of Iglehart and Jaquette. The proof is provided in the consecutive paper [6]. We developed also an approximation model which has emerged out of our short-term and long-term evaluation of the ordering procedures under consideration. For this purpose we devote a significant portion of the present study to the analysis of the results of several simulation runs, which have been designed to yield information on the rates at which the ordering policies lead to stable inventories, and to give clues on simple relations between the orders of the lower echelon stations and those of the upper echelon. Simple approximations to the upper echelon ordering levels in terms of those of the lower echelon stations have been established. We used also the simulated systems to investigate the question of what confidence levels γ to use in the Bayes prediction policy. The important result is that the average monthly cost is quite insensitive to changes in γ over a wide interval of γ values. As in the previous papers, we start with a discussion of the inventory model (sec. 1), proceed to discuss the cost function and the Bayesian framework (sec. 2). The algorithm for the determination of the optimal lower echelon orderings is presented in sec. 3. Section 4 presents the upper echelon's Bayes prediction policy. Section 5 is devoted to the development of a procedure which is based on the asymptotic properties of the adaptive procedure of the earlier sections. In sec. 6 we present and analyze the results of the simulations. Several conclusions are drawn concerning the properties of the adaptive procedure and its asymptotic counterpart. Section 7 provides derivations and proofs.

1. THE INVENTORY MODEL

Consider a two-echelon inventory system in which there are s stations at the lower echelon. These stations are labelled $E_1, \ldots, E_s$. The upper echelon is designated by D. Let $Q_{n,i}$ denote the stock level in station E_i $(i = 1, \ldots, s)$ at the beginning of the nth month. An order of size $Y_{n,i}$ is issued by station E_i at the beginning of the nth month. The order levels satisfy the inequalities

$$Y_{n,i} \geq -Q_{n,i}, \qquad i = 1, \ldots, s. \tag{1.1}$$

That is, $Y_{n,i}$ could be negative. A negative value of $Y_{n,i}$ signifies the shipment from E_i to D. Let S_n designate the stock level of the *whole system* at the beginning of the nth month, and let V_n designate the

stock level at D at the beginning of the nth month. That is,

$$V_n = S_n - \sum_{i=1}^{s} Q_{n,i}, \qquad n = 1, 2, \ldots. \tag{1.2}$$

$V_n \geq 0$ for all $n = 1, 2, \ldots$. Since the stations can obtain stock only from the depot, D, and replenishment has a leadtime of 1 month, we impose the condition

$$\sum_{i=1}^{s} Y_{n,i}^{+} \leq V_n, \tag{1.3}$$

where $Y_{n,i}^{+} = \max\{Y_{n,i}, 0\}$.

2. THE COST FUNCTION AND THE BAYESIAN FRAMEWORK

As in the previous study [7], we consider the inventory cost for the nth month to be based upon holding and shortage costs at each one of the s stations. This cost function is of the form

$$C(\underset{\sim}{Q}_n, \underset{\sim}{X}_n) = \sum_{i=1}^{s} \{c_i(Q_{n,i} - X_{n,i})^{+} + p_i(Q_{n,i} - X_{n,i})^{-}\}, \tag{2.1}$$

where $\underset{\sim}{Q}_n = (Q_{n,1}, \ldots, Q_{n,s})$, $a^{+} = \max(a, 0)$, $a^{-} = -\min(a, 0)$, and $X_{n,i}$ designates a random variable representing the demand at the ith station during the nth month. c_i is the carrying cost for an undemanded unit and p_i is the shortage cost, at the ith station. We assume that the demands at the various stations are independent random variables with Poisson distributions of means $\lambda_i (i = 1, \ldots, s)$. Let $T_{n,i} = \sum_{j=0}^{n} X_{j,i}$, $X_0 \equiv 0$. This is a sufficient statistic for λ_i. We consider a Bayesian framework in which $\lambda_1, \ldots, \lambda_s$ are priorly independent. The prior distribution of λ_i is the gamma distribution with scale parameters τ_i and shape parameters ν_i. This prior gamma has a density proportional to $\lambda^{\nu_i - 1} \exp\{-\lambda/\tau_i\}$. As previously shown [7], the anticipated distribution of $X_{n,i}$ given $T_{n-1,i}$ is the negative binomial N.B. $(\psi_{n,i}, \nu_i + T_{n-1,i})$, where $\psi_{n,i} = \tau_i/(1 + n\tau_i)$. An N.B. (ψ, ν) has a probability mass function

$$g(x|\psi, \nu) = \frac{\Gamma(\nu + x)}{\Gamma(\nu)\Gamma(x + 1)} \psi^{x}(1 - \psi)^{\nu}, \qquad x = 0, 1, \ldots \tag{2.2}$$

and a c.d.f. which we denote by $G(x|\psi, \nu)$. The anticipated inventory cost of the $(n + 1)$st month, given the total demand vector $\underset{\sim}{T}_{n-1}$ and the vector $\underset{\sim}{Y}_n$, is

$$R(\underset{\sim}{Y}_n; \underset{\sim}{Q}_n, \underset{\sim}{T}_{n-1}) = \sum_{i=1}^{s} E_{i,n}\{c_i[(Q_{n,i} + Y_{n,i} - X_{n,i})^{+} - X_{n+1,i}]^{+} + p_i[(Q_{n,i} + Y_{n,i} - X_{n,i})^{+} - X_{n+1,i}]^{-}\}; \tag{2.3}$$

where $E_{i,n}\{\cdot\}$ denotes the expectation with respect to the anticipated joint distribution of $(X_{n,i}, X_{n+1,i})$. As shown in [7], this anticipated joint distribution has a bivariate joint probability mass function

$$g(x, y|\tau_i, \nu_i+T_{n-1,i}) = g(x|\psi_{n,i}, \nu_i+T_{n-1,i})g(y|\psi_{n+1,i}, \nu_i+T_{n-1,i}+x) \tag{2.4}$$

for all $x, y=0, 1, 2, \ldots$. As shown in [2], the optimal ordering vector $\underset{\sim}{Y}_n^0$ is obtained by minimizing (2.3) under the constraints (1.1) and (1.3), irrespective of the ordering policy of the upper echelon. In the next section we discuss this minimization process.

3. AN ALGORITHM FOR THE DETERMINATION OF THE OPTIMAL LOWER ECHELON ORDERING LEVELS

In sec. 7.1 we show how to determine an unconstrained minimization of each term of (2.3). According to this unconstrained minimization procedure we determine for each station E_i the integer

$$k_{n,i} = \text{least nonnegative integer, } k, \text{ such that} \tag{3.1}$$

$$G(k|2\psi_{n+1,i}, \nu_i + T_{n-1,i}) \geq \frac{p_i}{c_i + p_i} G(k|\psi_{n,i}, \nu_i + T_{n-1,i}) .$$

The vector $\underset{\sim}{k}_n = (k_{n,1}, \ldots, k_{n,s})$ is the starting point in the search for the optimal vector $\underset{\sim}{Y}_n^0$. We present here the formal algorithm. The proof of its optimality is given in sec. 7.2.

An Algorithm for the Determination of $\underset{\sim}{Y}_n^0$

STEP 1. Set $A \leftarrow \{1, 2, \ldots, s\}$; $R_i \leftarrow \dfrac{p_i}{c_i + p_i}$, $(i = 1, \ldots, s)$;

$$\nu_{n,i} \leftarrow \nu_i + T_{n-1,i}, \qquad (i = 1, \ldots, s);$$

$$\psi_{n,i} \leftarrow \tau_i/(1+ n\tau_i) \text{ and } \psi_{n+1,i} \leftarrow \tau_i/(1 + (n+1)\tau_i), \qquad (i = 1, \ldots, s).$$

STEP 2. For each $i\epsilon A$ determine

$$k_{n,i} = \text{least nonnegative integer, } k, \text{ such that}$$

$$G(k|2\psi_{n+1,i}, \nu_{n,i}) \geq R_i G(k|\psi_{n,i}, \nu_{n,i});$$

and

$$d_{n,i} = k_{n,i} - Q_{n,i}.$$

STEP 3. If $\sum_{i\epsilon A} (k_{n,i} - d_{n,i})^+ \leq V_n$, set $Y_{n,i}^0 \leftarrow d_{n,i}$ and stop; else, go to Step 4.

STEP 4. For each $i\epsilon A$ set $Y_{n,i}^0 \leftarrow d_{n,i}$ and $A \leftarrow A - \{i\}$ if $d_{n,i} \leq 0$.

STEP 5. For each $i\epsilon A$ determine

$$D_i(k_{n,i}) = (c_i + p_i)G(k_{n,i} - 1|2\psi_{n+1,i}, \nu_{n,i}) - p_i G(k_{n,i} - 1|\psi_{n,i}, \nu_{n,i}).$$

STEP 6. Determine an integer j in A such that

$$|D_j(k_{n,j})| = \min_{i\epsilon A} |D_i(k_{n,i})|.$$

STEP 7. Set

$$k_{n,i} \leftarrow \begin{cases} k_{n,i}, & \text{if } i \neq j \\ k_{n,j}-1, & \text{if } i=j \end{cases}, \qquad i\epsilon A;$$

$$d_{n,i} \leftarrow k_{n,i} - Q_{n,i}, \qquad i\epsilon A.$$

Go to Step 3.

Notice that the algorithm checks first whether the unconstrained solution (the vector of "base" levels) satisfies the constraint. If it does, it is also a constrained solution. If it does not, the $Y_{n,i}$ values are reduced gradually; in each step a station is chosen so that the expected increase in the objective function is minimal. After each reduction the feasibility of the current $Y_{n,i}$ values is checked. Whenever a $Y_{n,i}$ value is not positive it is optimal, and the respective station does not take part in the continuing search. The search terminates as soon as the constraint is satisfied.

4. BAYES PREDICTION ORDERING POLICY FOR THE UPPER ECHELON

In the present section, we develop an ordering policy for the upper echelon, which is an s-station generalization of the policy discussed in [7]. The main idea is to determine an ordering policy which insures, with a preassigned posterior probability, γ, that the stock level at the upper echelon will be sufficient to provide the demand of the lower echelon stations without restriction and not accumulating too much inventory. γ is a specified prediction (Bayes) confidential level, $0 < \gamma < 1$.

We wish to find the least nonnegative integer, Z, such that

$$P\left\{Z + Z_{n-1} + V_n \geq \sum_{i=1}^{s} (Y_{n,i} + Y_{n+1,i}) + \sum_{i=1}^{s} (k_{n+2,i} - Q_{n+2,i}) | \underline{T}_{n-1}\right\} \geq \gamma, \tag{4.1}$$

where Z_{n-1} is the quantity ordered by the upper echelon at the beginning of the $(n-1)$st month. We denote by Z_n the quantity determined according to (4.1). The evaluation of Z_n according to (4.1) is not a simple matter, due to the statistical dependence of the $Y_{n,i}$ and the $Q_{n,i}$ variables. Moreover, whenever sufficient stock exists at the upper echelon one can substitute $k_{n,i} - Q_{n,i}$ for the optimum $Y^0_{n,i}$. We therefore consider an ordering policy which is based on the following modification of (4.1), namely

(4.2) Z_n = least nonnegative integer, Z, such that

$$P\left\{Z + V_{n+1} \geq \sum_{i=1}^{s} (H_{n+1,i} + H_{n+2,i}) | \underline{T}_{n-1}\right\} \geq \gamma,$$

where for each $i = 1, \ldots, s$

$$H_{n+j,i} = k_{n+j-1,i} - (k_{n+j-1,i} - X_{n+j-1,i})^+, \qquad j = 1, 2, \tag{4.3}$$

and $V_{n+1} = Z_{n-1} + V_n - \sum_{i=1}^{s} Y_{n,i}$. We notice that $H_{n+j,1}, \ldots, H_{n+j,s}$ are independent $(j=1,2)$. Hence, if we can determine for each i $(i=1, \ldots, s)$ the anticipated probability density of $H_{n+1,i} + H_{n+2,i}$ given $\underset{\sim}{T}_{n-1}$ then the anticipated probability distribution of $\sum_{i=1}^{s} (H_{n+1,i} + H_{n+2,i})$ can be easily determined by an s-fold convolution. Let

$$\psi_{n,i}(x) = P\{H_{n+1,i} + H_{n+2,i} = x | \underset{\sim}{T}_{n-1}\}. \qquad x = 0, 1, \ldots, \qquad i = 1, \ldots, s.$$

In sec. 7.3 we provide an explicit derivation of these functions. Here we continue with the required convolution. Let $\zeta_{n,s}(x)$ denote the anticipated distribution of $\sum_{i=1}^{s} (H_{n+1,i} + H_{n+2,i})$ given $\underset{\sim}{T}_{n-1}$, $\zeta_{n,s}(x)$ can be obtained according to the recursive equation

$$\begin{cases} \zeta_{n,1}(x) = \psi_{n,1}(x), & x = 0, 1, \ldots \\ \zeta_{n,j+1}(x) = \sum_{y=0}^{x} \zeta_{n,j}(x-y)\psi_{n,j+1}(y), & j = 1, \ldots, s-1. \end{cases} \tag{4.4}$$

After determining $\zeta_{n,s}(x)$ we can determine Z_n^* in the following manner: Let $\chi_{n,s}(\gamma)$ denote the γ-fractile of the anticipated distribution of $\sum_{i=1}^{s} (H_{n+1,i} + H_{n+2,i})$, given $\underset{\sim}{T}_{n-1}$; i.e.,

$$\chi_{n,s}(\gamma) = \text{least nonnegative integer, } j, \text{ such that} \tag{4.5}$$

$$\sum_{x=0}^{j} \zeta_{n,s}(x) \geq \gamma.$$

The value of Z_n is equal to $\left(\chi_{n,s}(\gamma) - \left(Z_{n-1} + V_n - \sum_{i=1}^{s} Y_{n,i}\right)\right)^+$

5. THE ASYMPTOTIC BEHAVIOR OF THE ORDERING POLICIES

In the present section we investigate the asymptotic behavior of the ordering policies as $n \to \infty$. In other words, what is the nature of the policies as time goes on, and the effect of the prior assumptions on the limiting behavior.

We start the investigation with the following observation. *If a sequence of i.i.d. random variables* $X_1, X_2, \ldots$ *has a common Poisson distribution with (unknown) mean* λ_0, $0 < \lambda_0 < \infty$. *If the unknown mean,* λ, *is assigned a prior gamma distribution* $G\left(\frac{1}{\tau}, \nu\right)$, *with any prior parameters* (ν, τ), $0 < \nu, \tau < \infty$, *then the anticipated distribution of* X_{n+1}, *given* n, $T_n = \sum_{i=1}^{n} X_i$ (*namely, the negative binomial N.B.*

(ψ_n, ν_n), $\psi_n = \tau/(1 + n\tau)$, $\nu_n = X + T_n$) *converges to the Poisson distribution* $P(\lambda_0)$. The proof of this assertion is given in sec. 7.4. Thus, we deduce from the present limiting result that all the anticipated distributions which play a role in the development of the ordering policies of the lower and the upper echelons converge (a.s.) to corresponding Poisson distributions, whose means are those according to which the demand variables are generated. This result can be used also to overcome a technical difficulty which arises whenever one has to compute the negative binomial distribution N.B. $(\psi_n, \nu+T_{n-1})$ for large values of n. In these cases, if the mean λ_0 of the parent Poisson distribution is not very small, we expect the values of T_{n-1} to be generally quite large. As a result, the usual computer routines for the calculation of the N.B. $(\psi_n, \nu+T_{n-1})$ probability distributions, and other related functions, become

TABLE 1. *The values of* k_n *and* $\hat{k}_n$ *for* $n = 1(1)50$, $\nu = 1$, $\tau = 1$, *and* $\lambda = 10$

n	k_n	$\hat{k}_n$	n	k_n	$\hat{k}_n$	n	k_n	$\hat{k}_n$	n	k_n	$\hat{k}_n$	n	k_n	$\hat{k}_n$
1	29	14	11	26	24	21	24	23	31	25	24	41	26	25
2	30	19	12	26	24	22	24	23	32	25	24	42	26	26
3	32	23	13	25	23	23	24	23	33	25	25	43	26	25
4	29	23	14	25	23	24	25	24	34	26	25	44	26	25
5	28	23	15	25	23	25	24	23	35	26	25	45	26	25
6	26	22	16	25	23	26	24	23	36	26	25	46	26	25
7	27	23	17	24	22	27	25	24	37	26	25	47	26	25
8	28	24	18	24	23	28	25	24	38	26	26	48	25	25
9	26	23	19	24	23	29	25	24	39	26	25	49	25	25
10	27	24	20	24	23	30	25	24	40	24	25	50	25	25

unreliable due to possible underflow and similar phenomena. One could overcome this difficulty by using instead Poisson distributions with means $\hat{\lambda}_n = \psi_n(\nu+T_{n-1}) = \tau(\nu+T_{n-1})/(1+n\tau)$. $\hat{\lambda}_n$ is the Bayes estimator of λ, with respect to a squared-error loss function. We notice that $\hat{\lambda}_n \to \lambda_0$ a.s. as $n \to \infty$, where λ_0 is the actual value of λ. In Table 1 we compare the base stock levels k_n defined in (3.1) (for simplicity we delete here the subscript i of the station) to corresponding values $\hat{k}_n$ defined in terms of Poisson distributions in the following manner:

(5.1) $\hat{k}_n$ = least nonnegative integer, k, such that

$$P(k; 2\hat{\lambda}_n) \geq R \cdot P(k; \hat{\lambda}_n),$$

where $R = p/(c + p)$ is the cost ratio, and $P(j; \theta)$ is the c.d.f. of a Poisson with mean θ. The values in Table 1 were computed according to (3.1) and (5.1) with the prior parameters $\nu = 1$ and $\tau = 1$, and the values of $X_1, X_2, \ldots$ were simulated according to a Poisson distribution with mean $\lambda_0 = 10$. As seen in this table, there is at the beginning a considerable discrepancy between the functions k_n and $\hat{k}_n$. However, for all $n \geq 10$ the two functions yield values which do not differ by more than 2 units. For large values of n the two functions yield the same values.

It is interesting to note in this connection that the computation of the 50 k_n values, together with the simulation of the demand values, consumed on a time-sharing GE-Honeywell computer 156 sec. However, the computations of the $\hat{k}_n$ values required only 16.5 sec. We expect that ordering levels which are computed on the basis of Poisson distributions will consume less computing time than those

based on the original negative binomial distributions. If one applies ordering policies based on Poisson distributions, the lower echelon ordering values $\hat{Y}_{n,i}$ can be determined according to the algorithm given in sec. 3, replacing the $k_{n,i}$ functions by the corresponding $\hat{k}_{n,i}$ functions, which are computed according to (5.1). The upper echelon ordering level Z_n^* is asymptotically (as n grows) equal to the value

$$\hat{Z}_n = \left(\hat{\chi}_n(\gamma) - \left(V_n + \hat{Z}_{n-1} - \sum_{i=1}^{s} \hat{Y}_{n,i}\right)\right)^+,$$

where $\hat{\chi}_n(\gamma)$ is the γ-fractile of the distribution of $\sum_{i=1}^{s} (\hat{H}_{n+1,i} + \hat{H}_{n+2,i})$. These $\hat{H}_{n+j,i}$ $(j=1, 2)$ functions are defined analogously to the H-functions given in (4.3), i.e.,

$$\hat{H}_{n+j,i} = \hat{k}_{n,i} - (\hat{k}_{n,i} - \hat{X}_{n+j-1,i})^+, \qquad j=1, 2. \tag{5.2}$$

In the present case $\hat{X}_{n,i}$ and $\hat{X}_{n+1,i}$ $(i=1, \ldots, s)$ are independent identically distributed Poisson random variables, with means $\hat{\lambda}_{n,i} = \psi_{n,i}(\nu_i + T_{n-1,i})$. As in sec. 4, the starting point is the determination of the probability density functions

$$\hat{\psi}_{n,i}(x) = P[\hat{H}_{n+1,i} + \hat{H}_{n+2,i} = x].$$

Under the above assumptions $\hat{H}_{n+1,i}$ and $\hat{H}_{n+2,i}$ are independent and identically distributed. Therefore, the probability density function $\hat{\psi}_{n,i}(x)$ is determined as a convolution of the density function

$$P[\hat{H}_{n+j,i}=x] = \begin{cases} p(x; \hat{\lambda}_{n,i}) & x \leq \hat{k}_{n,i} - 1 \\ 1 - P(k_{n,i} - 1; \hat{\lambda}_{n,i}), & x - k_{n,i} \\ 0, & \text{otherwise,} \end{cases} \tag{5.3}$$

with itself. $p(j; \theta)$ designates the probability density function of a Poisson distribution with mean θ. Accordingly, we obtain, for every $x=0, 1, \ldots,$

$$\hat{\psi}_{n,i}(x) = I\{x \leq 2\hat{k}_{n,i} - 2\} \sum_{y=(x-k_{n,i}+1)^+}^{\min\{x, k_{n,i}-1\}} p(y; \hat{\lambda}_{n,i}) p(x-y; \hat{\lambda}_{n,i}) \tag{5.4}$$

$$+ 2(1 - P(\hat{k}_{n,i} - 1; \hat{\lambda}_{n,i})) I\{\hat{k}_{n,i} \leq x \leq 2\hat{k}_{n,i} - 1\} p(x - \hat{k}_{n,i}; \hat{\lambda}_{n,i})$$

$$+ I\{x = 2\hat{k}_{n,i}\} (1 - P(\hat{k}_{n,i} - 1; \hat{\lambda}_{n,i}))^2.$$

The functions $\hat{\zeta}_{n,s}(x)$ and $\chi_{n,s}(\gamma)$ are defined in analogy to (4.4) and (4.5), respectively.

6. MONTE CARLO SIMULATION

6.1. A Six-Station System

We start this section with a discussion of the characteristics revealed by a simulated system of six stations in the lower echelon. The Monte Carlo experiment under consideration will be labelled

Case A. It is comprised of the following parameters:

System parameters in Case A

Station (i)	ν_i	τ_i	Q_i	c_i	p_i	λ_i	V	Z
1	19.	2.	0	2.	12.	2.	–	–
2	19.	2.	2	4.	10.	4.	–	–
3	19.	2.	4	6.	8.	6.	–	–
4	19.	2.	6	6.	8.	8.	–	–
5	19.	2.	8	4.	10.	10.	–	–
6	19.	2.	10	2.	12.	12.	–	–
Depot	–	–	–	–	–	–	1	0

The equal values of the prior parameters ν and τ reflect a situation in which the controller expects that the actual means λ_i of the various Poisson distributions of demand are equal. The cost structure is such that we can study the influence of different values of λ on the ordering policies at the same cost ratio. In Table 2 we present the simulated values $Q_{n,i}(i=1,\ldots,6)$, $Y_{n,i}\,(i=1,\ldots,6)$, V_n, Z_n,

TABLE 2. *Simulated stock levels at six stations, depot, ordering levels and monthly costs, according to procedures I and II, for 24 months; parameters of Case A, and* $\gamma=0.9$

n	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	V	Z	CST
1	0	2	4	6	8	10	5	0	0	0	0	0	5	62	108
	0	2	4	6	8	10	0	0	0	0	0	5	5	95	164
2	0	0	0	1	0	0	0	0	0	0	0	0	0	1	314
	0	0	0	1	0	4	0	0	0	0	0	0	0	0	266
3	0	0	0	0	0	0	10	10	7	10	11	14	62	63	94
	0	0	0	0	0	0	6	9	12	17	22	29	95	95	186
4	9	6	2	1	0	0	0	0	0	0	0	7	7	12	312
	5	5	7	8	10	13	0	0	0	0	0	0	0	0	116
5	9	0	0	0	0	0	0	10	9	10	15	19	63	66	86
	5	0	0	5	0	0	1	9	12	12	22	29	95	85	142
6	6	2	0	1	3	4	0	0	0	0	0	12	12	19	252
	3	1	0	8	10	14	0	0	0	0	3	7	10	10	208
7	4	0	0	0	0	4	3	10	10	12	15	16	66	68	120
	1	0	0	0	7	9	5	9	12	17	15	20	85	78	202
8	5	7	2	4	7	9	0	1	0	3	6	9	19	19	98
	4	6	4	9	14	18	0	0	0	5	7	5	17	17	156
9	3	4	0	0	2	14	5	8	13	15	16	8	68	65	202
	2	2	0	7	10	19	4	7	12	10	12	10	78	55	222
10	7	9	11	4	10	11	0	0	0	9	5	8	22	21	252
	5	6	10	6	14	18	1	3	2	11	8	11	40	36	228
11	6	4	9	5	7	6	2	7	2	10	11	17	65	49	200
	5	4	10	9	14	16	1	5	2	8	8	13	59	37	220
12	5	9	0	8	8	17	2	2	12	7	9	5	37	38	156
	3	7	0	10	12	23	3	2	12	7	10	6	58	40	186
13	5	10	7	3	8	7	3	0	6	13	10	17	49	50	200
	4	8	7	4	13	14	2	1	5	13	9	15	55	45	218
14	6	9	3	9	15	10	2	1	10	7	2	14	38	36	154
	4	8	2	10	19	15	2	1	10	7	3	14	50	37	170
15	8	3	7	8	6	6	−1	7	6	8	12	19	52	52	154
	6	2	5	9	11	11	0	7	7	8	11	18	58	51	170
16	6	9	7	5	4	8	0	0	3	8	12	14	37	36	152

TABLE 2. *Simulated stock levels at six stations, depot, ordering levels and monthly costs, according to procedures I and II, for 24 months; parameters of Case A, and* $\gamma = 0.9$ – Continued

n	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	V	Z	CST
	5	8	6	6	8	12	0	0	6	11	14	13	44	44	148
17	4	3	2	6	0	12	1	6	9	8	17	11	52	52	102
	3	2	4	10	5	15	0	7	8	7	17	12	51	51	230
18	2	6	8	6	10	15	4	3	4	9	8	8	36	36	182
	0	6	9	9	15	19	6	3	3	8	7	10	44	37	192
19	0	4	5	6	11	11	7	6	7	10	8	14	52	52	148
	0	5	5	8	15	17	6	4	7	9	7	12	58	45	182
20	6	5	7	8	2	16	0	3	4	6	16	7	36	36	162
	5	4	7	9	5	20	1	5	5	8	17	9	50	45	210
21	4	5	3	8	9	9	3	5	10	8	10	16	52	52	148
	4	6	4	12	13	15	2	3	8	5	9	14	50	41	228
22	4	7	6	13	9	14	3	2	7	2	11	11	36	36	214
	3	6	5	14	12	18	3	3	7	3	10	11	54	37	206
23	6	4	7	9	8	15	1	6	6	7	12	11	62	43	186
	5	4	6	11	10	19	1	5	6	6	12	10	58	40	230
24	6	6	7	10	16	10	1	4	6	5	4	17	45	37	222
	5	5	6	11	18	13	1	4	6	6	4	16	55	37	188

and the actual monthly cost (CST), under *two* different policies, for 24 months. The values on the first line of each month were obtained by the Bayes ordering policies of secs. 3 and 4, with $\gamma = 0.9$. In our computations we actually assumed that prior to the first month of control there were 4 months of observations, and therefore all the ψ-values are actually ψ_{n+4}. Such an assumption may effect the results only at the beginning. The values on the second line of each month represent the asymptotic ordering policy, which is described in sec. 5, based on the actual values of λ and $\gamma = 0.9$.

The comparison of the first and the second line of each month provides information on what can be done if the actual values of λ are known. As expected, the Bayes adaptive policy yields values of $Y_{n,i}$ and Z_n which converge to the asymptotic values as n grows. As seen in Table 2, in 17 out of the 24 months the simulated cost of the adaptive procedure is smaller than that of the asymptotic procedure. This phenomenon is illustrated in Table 3, in which the moving averages of monthly costs are presented.

TABLE 3. *Moving averages of the monthly costs in Case A, for the adaptive and the asymptotic procedures*

n	Moving averages		n	Moving averages	
	Adaptive	Asymptotic		Adaptive	Asymptotic
1	108.00	164.00	13	184.15	193.38
2	211.00	215.00	14	182.00	191.71
3	172.00	205.33	15	180.13	190.27
4	207.00	183.00	16	178.37	187.62
5	182.80	174.80	17	173.88	190.12
6	194.33	180.33	18	174.33	190.22
7	183.71	183.43	19	172.95	189.79
8	173.00	180.00	20	172.40	190.80
9	176.22	184.67	21	171.24	192.57
10	183.80	189.00	22	173.18	193.18
11	185.27	191.82	23	173.74	194.78
12	182.83	191.33	24	175.75	194.50

The moving averages of the monthly costs corresponding to the two procedures are close at the beginning, but since the eighth month they are considerably smaller under the adaptive procedure. The only reasonable explanation for this phenomenon is that although the asymptotic procedure should provide on the average smaller monthly costs, the adaptive procedure relies on the actual demand, especially during the first few months, and it can therefore yield sometimes better results than the asymptotic procedure.

An important characteristic of the ordering policy of the upper echelon is nicely revealed in Table 2. If we consider the $\hat{Z}_n$ values of the asymptotic procedure we realize that, for every $n \geq 3$, the value of $\hat{Z}_n$ equals exactly the sum of the corresponding $Y_{n,i}$ values. Indeed, the asymptotic procedure is based on a constant value of the γ-fractile of the distribution of $\sum_{i=1}^{s} (\hat{H}_{n+1,i} + \hat{H}_{n+2,i})$ say χ, which does not actually depend on n. We therefore obtain from the recursive equations

$$\text{(6.1)} \qquad \begin{cases} \hat{Z}_n = \left(\chi - V_n - \hat{Z}_{n-1} + \sum_{i=1}^{s} \hat{Y}_{n,i} \right)^+ \\ \qquad\qquad\qquad\qquad\qquad\qquad n = 2, 3, \ldots, \\ V_n = V_{n-1} + \hat{Z}_{n-2} - \sum_{i=1}^{s} \hat{Y}_{n-1,i} \end{cases}$$

that for each $n = 1, 2, \ldots$

$$\text{(6.2)} \qquad \hat{Z}_n = \begin{cases} \sum_{i=1}^{s} \hat{Y}_{n,i}, & \text{if } V_n > 0,\ \hat{Z}_{n-1} > 0 \\ V_{n-1} + \hat{Z}_{n-2} - \sum_{i=1}^{s} \hat{Y}_{n-1,i}, & \text{if } V_n = 0,\ \hat{Z}_{n-1} > 0 \\ \left(\chi - V_{n-1} - \hat{Z}_{n-2} + \sum_{i=1}^{s} (\hat{Y}_{n-1,i} + \hat{Y}_{n,i}) \right)^+, & \text{if } V_n > 0,\ \hat{Z}_{n-1} = 0 \\ \chi + \sum_{i=1}^{s} \hat{Y}_{n,i}, & \text{if } V_n = 0,\ \hat{Z}_{n-1} = 0. \end{cases}$$

The important feature of $\hat{Z}_n$, as presented by (6.2), is that whenever $\hat{Z}_{n-1} > 0$ it does not depend on χ. As seen in Table 2, all the values of Z_n of the *adaptive* procedure are greater than zero. The procedure by which these values were obtained is quite complicated (as described in sec. 4) and requires heavy and time consuming computations. It is therefore interesting to determine whether the Z_n values of the adaptive procedure can be approximated for every $n \geq 2$ by a simple rule, as the one given by (6.2). In order to obtain an answer to this question we computed the coefficient of correlation between the actual values of Z_n, yielded by the Bayes adaptive procedure, and the values $\hat{Z}_n$ obtained from (6.2) by using the V_n and $Y_{n,i}$ values of the adaptive procedure. These correlations are presented in Table 4. In this table we present 16 correlations. The coefficient of correlation ρ_j $(j = 1, \ldots, 16)$ is computed on the basis of $45 - j + 1$ pairs $\{(Z_n, \hat{Z}_n): j \leq n \leq 45\}$. (Case A was simulated with the adaptive pro-

cedure for 60 months.) We expect the approximation to Z_n which is provided by $\hat{Z}_n$ to be better as n grows. Thus, by eliminating the initial points, $(Z_n, \hat{Z}_n)$ we expect the correlation to get closer to 1. Once the value 1 is attained it never decreases by further elimination of points.

TABLE 4. *Correlations between* Z_n *and* $\hat{Z}_n$ *in case A under the adaptive procedure the correlation* ρ_j *is based on* $\{(Z_n, \hat{Z}_n) : j \leq n \leq 45\}$

j	1	2	3	4	5	6	7	8
ρ_j	0.993	0.993	0.994	0.999	0.999	0.999	0.999	0.999
j	9	10	11	12	13	14	15	16
ρ_j	0.993	0.998	0.998	0.999	0.999	0.999	0.999	1.000

In Table 4 we present only 16 correlations because the 16th one is already equal to 1. The correlations in Table 4 reveal that the values of Z_n can be determined almost perfectly by the corresponding $\hat{Z}_n$ values. *The conclusion from these trials is that it is sufficient to determine the actual* Z_n *values, according to the Bayes adaptive procedure, only at the first few months of the control. Then one could switch to an upper echelon ordering policy which is based on* $\hat{Z}_n$. These values would generally be equal to the sums of the corresponding $Y_{n,i}$ values. Such a modification of the upper echelon's ordering policy would generally result in a substantial saving of computing time.

6.2. A Three-Station System

We present now the simulation results of another case, Case B. In this case we simulated a system with three lower echelon stations. The three stations have the same parameters, which are: $\lambda = 5$, $Q = 5$, $v = 10$, $\tau = 2$, and cost parameters $c = \$1$ and $p = \$7$. We performed three runs (with the same simulated random demand values). These three runs differ with respect to the initial stock level at the upper echelon, V_1, and with respect to the value Z_0 ordered before the control started. These values are:

Upper echelon parameters in Case B

Run 1: $V_1 = 9$, $Z_0 = 0$
Run 2: $V_1 = 100$, $Z_0 = 0$
Run 3: $V_1 = 9$, $Z_0 = 60$.

The simulated values of V_n, $Q_{n,i}$, $Y_{n,i}$, and Z_n for 12 months ($n = 1, \ldots, 12$) are presented for each run in Table 5. We see in this table that when V_1 is too small and $Z_0 = 0$ (Run 1) there are considerable fluctuations in the V_n and Z_n values during the first 12 months. These fluctuations cause generally undesired losses due to shortages in the lower echelon stations. On the other hand, if V_1 is large (Run 2) then even if $Z_0 = 0$ there will be no orders issued by the upper echelon during the first few months. The system operates with no constraints and reaches stability very soon ($n \geq 5$). The comparison of Run 3 with Run 2 shows that if one starts to control a system which is understocked, an initial high order ($Z_0 = 60$) will bring the system into a stable phase in a short period of time.

TABLE 5. *Simulated stock levels at three stations and a depot, Case B*

	n	V_n	$X_{n,1}$	$X_{n,2}$	$X_{n,3}$	$Q_{n,1}$	$Q_{n,2}$	$Q_{n,3}$	$Y_{n,1}$	$Y_{n,2}$	$Y_{n,3}$	Z_n
	1	9	5	8	4	5	5	5	3	4	2	29
	2	0	4	5	10	3	1	3	0	0	0	3
	3	29	7	4	8	0	0	0	9	9	11	31
	4	3	5	3	5	2	5	3	0	0	3	3
	5	31	9	4	3	0	2	1	12	8	11	32
Run 1	6	3	8	3	8	3	6	9	3	0	0	4
	7	32	4	7	8	0	3	1	12	8	12	33
	8	4	3	2	3	8	4	5	0	0	4	2
	9	33	6	7	3	5	2	6	9	10	8	28
	10	8	3	9	6	8	5	11	3	5	0	8
	11	28	6	9	6	8	1	5	5	13	9	28
	12	9	7	4	8	7	5	8	3	4	2	10
	1	100	5	8	4	5	5	5	5	7	5	0
	2	83	4	5	10	5	4	6	5	8	7	0
	3	63	7	4	8	6	7	3	6	5	11	0
	4	41	5	3	5	5	8	6	7	4	8	12
	5	22	9	4	3	7	9	9	6	3	4	14
Run 2	6	21	8	3	8	4	8	10	10	3	4	18
	7	18	4	7	8	6	8	6	7	3	8	19
	8	18	3	2	3	9	4	6	4	7	7	16
	9	19	6	7	3	10	9	10	4	3	4	12
	10	24	3	9	6	8	5	11	5	8	3	16
	11	20	6	9	6	10	4	8	3	10	6	20
	12	17	7	4	8	7	5	8	5	7	5	18
	1	9	5	8	4	5	5	5	3	4	2	0
	2	60	4	5	10	3	1	3	7	11	10	0
	3	32	7	4	8	6	7	3	6	5	11	24
	4	10	5	3	5	5	8	6	4	1	5	10
	5	24	9	4	3	4	6	6	9	6	7	23
Run 3	6	12	8	3	8	4	8	10	8	2	2	13
	7	23	4	7	8	4	7	4	9	4	10	24
	8	13	3	2	3	9	4	6	2	5	6	11
	9	24	6	7	3	8	7	9	6	5	5	17
	10	19	3	9	6	8	5	11	5	8	3	16
	11	20	6	9	6	10	4	8	3	10	6	20
	12	17	7	4	8	7	5	8	5	7	5	18

6.3. A Two-Station System

In the present section we discuss the simulation results on a two-station system, which are designed to study the effect of the prediction threshold probability γ. In Table 6 we present the simulation results of a 29-month run, in which γ assumed the values 0.5, 0.6, 0.75 and 0.95. We label this run as Case C. The parameters of this case are:

$$\lambda_1 = \lambda_2 = 5, \qquad Q_1 = Q_2 = 1, \qquad V_1 = 9, \qquad Z_0 = 0, \qquad \nu_1 = 9, \quad \nu_2 = 10,$$

$$\tau_1 = 2, \qquad \tau_2 = 1, \qquad c_1 = c_2 = \$1, \qquad p_1 = p_2 = \$7.$$

TABLE 6. *A two-station simulation run under Case C.*[a] *Bayes adaptive policy*

n	$\gamma = 0.5$			$\gamma = 0.6$			$\gamma = 0.75$			$\gamma = 0.95$		
	V_n	Z_n	CST	V_n	Z_n	CST	V_n	Z_n	CST	V_n	Z_n	CST
1	9	13	2.00	9	14	2.00	9	16	2.00	9	21	2.00
2	0	1	63.00	0	1	63.00	0	1	63.00	0	1	63.00
3	13	14	5.00	14	15	6.00	16	17	8.00	21	22	13.00
4	1	1	49.00	1	1	42.00	1	1	36.00	1	1	25.00
5	14	15	3.00	15	17	4.00	17	19	7.00	22	22	13.00
6	1	2	49.00	1	1	42.00	1	1	21.00	3	3	5.00
7	15	15	10.00	17	18	12.00	19	20	14.00	22	22	21.00
8	2	2	15.00	1	0	8.00	1	0	3.00	4	2	1.00
9	15	16	20.00	18	19	7.00	20	21	10.00	23	15	13.00
10	2	2	57.00	0	0	64.00	0	0	59.00	11	11	9.00
11	16	17	6.00	19	20	9.00	21	22	13.00	15	16	13.00
12	2	2	0.00	0	0	9.00	0	0	5.00	11	11	16.00
13	17	17	10.00	20	20	15.00	22	22	20.00	16	11	20.00
14	2	2	6.00	0	0	9.00	0	0	15.00	16	7	21.00
15	17	16	12.00	20	16	15.00	22	11	15.00	20	4	15.00
16	2	2	45.00	3	3	33.00	10	10	8.00	22	11	9.00
17	16	17	7.00	16	17	9.00	11	12	7.00	15	16	12.00
18	2	2	28.00	3	3	7.00	10	10	4.00	11	11	10.00
19	17	18	9.00	17	18	9.00	12	13	8.00	16	17	18.00
20	2	1	3.00	3	3	4.00	10	10	10.00	11	10	20.00
21	18	18	9.00	18	17	10.00	13	12	11.00	18	6	15.00
22	1	1	24.00	3	4	3.00	10	11	11.00	21	13	17.00
23	18	18	10.00	17	17	9.00	12	12	12.00	15	10	16.00
24	1	2	11.00	4	4	5.00	11	11	15.00	18	11	19.00
25	18	17	10.00	17	17	10.00	12	12	15.00	17	7	15.00
26	2	3	3.00	4	4	5.00	11	11	17.00	20	13	18.00
27	17	17	6.00	17	17	8.00	12	10	13.00	15	9	13.00
28	3	3	36.00	4	4	23.00	13	13	13.00	19	14	14.00
29	17	17	8.00	17	17	9.00	10	10	13.00	14	14	18.00
			17.79			15.55			15.38			16.35

[a] CST designates the sum of the monthly cost of the two stations. The figures in the last row are the 29-month averages of the system cost.

As shown in Table 6, the 29-month cost average corresponding to $\gamma = 0.75$ is the smallest. However, the differences between the average costs are quite small. An interesting phenomenon revealed in Table 6 is concerned with the effect of γ on the fluctuations of the Z_n values. When $\gamma = 0.5$ or 0.6 the Z_n values fluctuate throughout the 29-month period. When $\gamma = 0.7$ the Z_n values reach a stable phase after 15 months. When $\gamma = 0.95$ the stable phase is attained after 9 months. For this reason the 12-month cost average for $\gamma = 0.95$ is \$16.17 against a cost average of \$17.58 for $\gamma = 0.75$. It seems that if one is interested in attaining the stability phase relatively fast the γ value to be employed should be around $\gamma = 0.90$.

7. PROOFS AND DERIVATIONS

7.1. The Base Stock Level of Lower Echelon Stations

In the present section we prove that the unconstrained minimum of the anticipated cost (2.3) is attained by the $k_{n,i}$ values, which are defined in (3.1). In order to simplify the notation we delete the

station subscript, i, and let k^0 designate the value of $k_{n,i}$. We have to prove that k^0 minimizes

$$R_n(k) = E\{c[(k - X_n)^+ - X_{n+1}]^+ + p[(k - X_n)^+ - X_{n+1}]^-\}, \tag{7.1}$$

where the expectation is with respect to the joint anticipated distribution of (X_n, X_{n+1}) given $\underset{\sim}{T}_n$. Let $\underset{k}{\Delta} R_n(k) = R_n(k+1) - R_n(k)$. As proven in [5],

$$\underset{k}{\Delta} R_n(k) = (c+p)G(k|2\psi_{n+1}, \nu_n) - pG(k|\psi_n, \nu_n), \tag{7.2}$$

where $\nu_n = \nu + T_{n-1}$. It is simple to verify that $2\psi_{n+1} > \psi_n$ for *all* τ. Moreover, ν_n is fixed for a given n. Since the family of negative binomial distributions with a fixed ν_n, and $0 < \psi < 1$, is a *monotone likelihood ratio* family (see Zacks [4], p. 431),

$$G(k|2\psi_{n-1}, \nu) \leq G(k|\psi_n, \nu), \qquad \text{for } \textit{all } k = 0, 1, \ldots; \tag{7.3}$$

However,

$$G(k|2\psi_{n+1}, \nu)/G(k|\psi_n, \nu) \to 1 \text{ as } k \to \infty.$$

Therefore, since $p/(c+p) < 1$, there exists a k^*, sufficiently large, such that $\underset{k}{\Delta} R_n(k^*) \geq 0$.

Let

$$k^0 = \text{least } k \geq 0 \text{ such that } \underset{k}{\Delta} R_n(k^0) \geq 0. \tag{7.4}$$

We prove now that $\underset{k}{\Delta^2} R_n(k) = R_n(k+2) - 2R_n(k+1) + R_n(k)$ is nonnegative for *all* $k \geq k^0 - 1$. This implies immediately that k^0 is optimal. Since (3.1) and (7.4) are equivalent, this will prove the optimality of (3.1).

Consider the second order difference. According to (7.2),

$$\begin{aligned} \underset{k}{\Delta^2} R_n(k) &= \underset{k}{\Delta}\{(c+p)G(k|2\psi_{n+1}, \nu_n) - pG(k|\psi_n, \nu_n)\} \\ &= (c+p)g(k+1|2\psi_{n+1}, \nu_n) - pg(k+1|\psi_n, \nu_n), \end{aligned} \tag{7.5}$$

where $g(k|\psi, \nu)$ is the probability function of the corresponding negative binomial distribution. Thus, $\underset{k}{\Delta^2} R_n(k) \geq 0$ if, and only if,

$$\frac{g(k+1|2\psi_{n+1}, \nu_n)}{g(k+1|\psi_n, \nu_n)} \geq \frac{p}{c+p}. \tag{7.6}$$

But,

$$\frac{g(k+1|2\psi_{n+1}, \nu_n)}{g(k+1|\psi_n, \nu_n)} = \left(\frac{2\psi_{n+1}}{\psi_n}\right)^{k+1} \left(\frac{1-2\psi_{n+1}}{1-\psi_n}\right)^{\nu_n}. \tag{7.7}$$

Let

$$\tilde{k} = \begin{cases} \text{least } k \geq 0 \text{ such that } \underset{k}{\Delta^2} R_n(k) \geq 0, \\ \infty, \text{ if no such } k \text{ exist.} \end{cases} \tag{7.8}$$

We notice first that if $\tilde{k} < \infty$ then $\underset{k}{\Delta^2} R_n(k) \geq 0$ for *all* $k \geq \tilde{k}$. Indeed, $2\psi_{n+1}/\psi_n > 1$; therefore, if (7.6) holds for some k_1 then it holds for all $k \geq k_1$. We prove now that $\tilde{k} \leq k^0 - 1$. Be definition of k^0,

$$R_n(k^0+1) \geq R_n(k^0).$$

Consider $\underset{k}{\Delta^2} R_n(k^0-1)$. Since $\underset{k}{\Delta^2} R_n(k^0-1) = R_n(k^0+1) - 2R_n(k^0) + R_n(k^0-1)$, and since

$$R_n(k^0) - R_n(k^0-1) < 0, \quad \underset{k}{\Delta^2} R_n(k^0-1) > 0.$$

It implies that $k^0 - 1 \geq \tilde{k}$, and the proof is complete.

7.2. Optimality of the Allocation Algorithm

We prove now that the algorithm given in sec. 3 minimizes (2.3) subject to the constratints (1.1) and (1.3). Starting with the unconstrained optimization given by the vector $\underset{\sim}{k}_n$, if $\sum_{i \in A} (k_{n,i} - Q_{n,i})^+ \leq V_n$ then (1.3) is satisfied, and by setting $Y^0_{n,i} = d_{n,i}$ we satisfy (1.1). This yields the minimal expected cost, as desired. If (1.3) is not satisfied we proceed to Step 4, in which we check for each station individually whether $d_{n,i} \leq 0$. We prove now that in all cases where $d_{n,i} \leq 0$ if we set $Y^0_{n,i} = d_{n,i}$ and delete the station from the optimization process, we act in an optimal manner. For this purpose, let $B = \{i;\ d_{n,i} \leq 0\}$ and $A = \{1, 2, \ldots, s\} - B$. Let $\underset{\sim}{Y}^A_n$ designate the subvector of $\underset{\sim}{Y}_n$ which consists of those components with indices in A, and let $\underset{\sim}{Y}^B_n$ be the complementary subvector. Let $R_1(\underset{\sim}{Y}^A_n; \cdot, \cdot) = \sum_{i \in A} E_{n,i}\{\cdot\}$, and $R_2(\underset{\sim}{Y}^B_n; \cdot, \cdot) = \sum_{i \in B} E_{n,i}\{\cdot\}$. Then $R(\underset{\sim}{Y}_n; \cdot, \cdot) = R_1(\underset{\sim}{Y}^A_n; \cdot, \cdot) + R_2(\underset{\sim}{Y}^B_n; \cdot, \cdot)$ and

$$\inf_{\underset{\sim}{Y}_n} R(\underset{\sim}{Y}_n; \cdot, \cdot) \geq \inf_{\underset{\sim}{Y}^A_n} R_1(\underset{\sim}{Y}^A_n; \cdot, \cdot) + \inf_{\underset{\sim}{Y}^B_n} R_2(\underset{\sim}{Y}^B_n; \cdot, \cdot). \tag{7.9}$$

Let $\underset{\sim}{\eta}^B_n = (Y^0_{n,i};\ i \epsilon B)$, where $Y^0_{n,i} = k_{n,i} - Q_{n,i}$, for each $i \epsilon B$, and let $\underset{\sim}{Y}^*_n = (\underset{\sim}{Y}^A_n, \underset{\sim}{\eta}^B_n)$. Since every component of $\underset{\sim}{\eta}^B_n$ is smaller or equal to zero, $\underset{\sim}{Y}^*_n$ satisfies the constraints (1.1) and (1.3) if all the components

of $\underset{\sim}{Y}_n^A$ satisfy these constraints. Furthermore, according to the definition of

$$k_{n,i}, R_2(\underset{\sim}{\eta}_n^B; \cdot, \cdot) = \inf_{\underset{\sim}{Y}_n^B} R_2(\underset{\sim}{Y}_n^B; \cdot, \cdot).$$

Finally, from the inequalities

$$\begin{aligned} \text{(7.10)} \qquad \inf_{\underset{\sim}{Y}_n^*} R(\underset{\sim}{Y}_n^*; \cdot, \cdot) &\geq \inf_{\underset{\sim}{Y}_n^*} R(\underset{\sim}{Y}_n; \cdot, \cdot) \\ &\geq \inf_{\underset{\sim}{Y}_n} R_1(\underset{\sim}{Y}_n^A; \cdot, \cdot) + R_2(\underset{\sim}{\eta}_n^B; \cdot, \cdot) \\ &= \inf_{\underset{\sim}{Y}_n^A} R(\underset{\sim}{Y}_n^*; \cdot, \cdot), \end{aligned}$$

the assertion is proven.

We show now that Steps 5–7 lead to the constrained minimization of $R_1(\underset{\sim}{Y}_n^A; \cdot, \cdot)$. Since $R_1(\underset{\sim}{Y}_n^A; \cdot, \cdot)$ is a sum of all the $E_{n,i}\{\cdot\}$, $i \epsilon A$, which are nonnegative; and since $k_{n,i}$ is determined to minimize $E_{n,i}\{\cdot\}$ the optimal adjustment should be in the direction of diminishing the $k_{n,i}$ values. Indeed, any increase in the values of $k_{n,i}$ may only result in an increase in the corresponding $E_{n,i}\{\cdot\}$ values. It is apparent that in each iteration one should choose to decrease that $k_{n,i}$ which yields the least increase in the associated $E_{n,i}\{\cdot\}$. After every such trial one should test for feasibility, terminating the process as soon as (1.3) is satisfied.

The reader may try to prove the optimality of the allocation algorithm by referring to a general theorem of Everett [1]. However, the specific development should be carried out along the lines of the present proof.

7.3. Derivation of the p.d.f. $\psi_{n,i}(x)$

We derive now the $\psi_{n,i}(x)$ functions. In order to understand the development of the formula of $\psi_{n,i}(x)$ it is worthwhile to write first the anticipated probability distribution of $H_{n+2,i}$ given $\underset{\sim}{T}_{n-1}$. According to the above remarks,

$$\begin{aligned} \text{(7.11)} \qquad P[H_{n+2,i}=x|\underset{\sim}{T}_{n-1}] &= I\{x < k_{n+1,i}\} g(x|\psi_{n+1,i}, \nu_{n+1,i}) \\ &\quad + I\{x = k_{n+1,i}\}(1 - G(k_{n+1,i}-1|\psi_{n+1,i}, \nu_{n+1,i})). \end{aligned}$$

We notice that $k_{n+1,i}$ and $\nu_{n+1,i}$ depend on the value of $X_{n,i}$. Hence, from (7.11),

$$\begin{aligned} \text{(7.12)} \qquad P[H_{n+1,i}+H_{n+2,i}=x|\underset{\sim}{T}_{n-1}, X_{n,i}=y] &= I\{x < k_{n+1,i}+m_i(y)\} g(x-m_i(y)|\psi_{n+1,i}\nu_{n,i}+Y) \\ &\quad + I\{x = k_{n+1,i}+m_i(y)\}[1 - G(k_{n+1,i}-1|\psi_{n+1,i}, \nu_{n,i}+y)], \end{aligned}$$

where $m_i(y) = \min(y, k_{n,i})$. Finally, $\psi_{n,i}(x)$ is the expectation of (7.12) with respect to the anticipated distribution of $X_{n,i}$, given $\underset{\sim}{T}_{n-1}$. Since this anticipated distribution is the N.B. $(\psi_{n,i}, \nu_{n,i})$ we obtain

the general formula

$$\psi_{n,i}(x)=\sum_{y=0}^{\infty} g(y|\psi_{n,i},\nu_{n,i})\left[\sum_{d=0}^{k_{n+1,i}-1} I\{x=m_i(y)+d\}\right.$$

$$\left.\cdot\, g(d|\psi_{n+1,i},\nu_{n,i}+y)+I\{x \quad =m_i(y)+k_{n+1,i}\}\cdot(1-G(k_{n+1,i}-1|\psi_{n+1,i} \quad ,\nu_{n,i}+y))\right]. \tag{7.13}$$

7.4. The Convergence of Posterior Distributions

According to the assumptions of sec. 2, the posterior distribution of λ given (n, T_{n-1}) is the gamma distribution $G(1/\psi_n, \nu_n)$. Accordingly, if $g(i|\psi_n, \nu_n)$ designates the probability density function of the anticipated distribution N.B. (ψ_n, ν_n), then by the Lebesgue Dominated Convergence Theorem

$$\lim_{n\to\infty} g(i|\psi_n,\nu_n)=\lim_{n\to\infty}\int_0^{\infty} p(i;\lambda)dH(\lambda|1/\psi_n,\nu_n)=\int_0^{\infty} p(i;\lambda)dH^*(\lambda), \tag{7.14}$$

where $p(i;\lambda)$ is a probability density function of a Poisson distribution $P(\lambda)$, and $H(\lambda|1/\psi_n, \nu_n)$ is the gamma c.d.f., with scale parameter ψ_n and shape parameter ν_n. Furthermore, $H^*(\lambda)$ designates the limiting distribution of $H(\lambda|1/\psi_n, \nu_n)$ as $n\to\infty$. We notice that the mean and the variance of $H(\lambda|1/\psi_n, \nu_n)$ are, respectively, $\psi_n\nu_n$ and $\psi_n^2\nu_n$. But

$$\lim_{n\to\infty}\psi_n\nu_n=\lim_{n\to\infty}\frac{\tau}{1+n\tau}(\nu+T_n)=\lambda_0 \qquad \text{a.s.}$$

$$\lim_{n\to\infty}\psi_n^2\nu_n=\lim_{n\to\infty}\left(\frac{\tau}{1+n\tau}\right)^2(\nu+T_n)=0 \qquad \text{a.s.} \tag{7.15}$$

Hence, $H^*(\lambda)$ is a one-point distribution concentrated at $\lambda=\lambda_0$. In the same manner we show that for any positive constant c,

$$\lim_{n\to\infty} g(i|c\psi_n,\nu_n)=p(i;c\lambda^0) \qquad \text{a.s.} \tag{7.16}$$

REFERENCES

[1] Everett, H., III, "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," Operations Research *11*, 399–417 (1963).

[2] Iglehart, D. L. and S. C. Jaquette, "Optimal Policies under the Shortage Probability Criterion for an Inventory Model with Unknown Dependent Demand," Nav. Res. Log. Quart. *16*, 485–493 (1969).

[3] Zacks, S., Bayes Adaptive Control of Two-Echelon Multistation Inventory Systems. Technical Memorandum Serial TM–61541, Institute for Management Science and Engineering, The George Washington University (1970).

[4] Zacks, S., *The Theory of Statistical Inference* (John Wiley & Sons, New York, 1971).

[5] Zacks, S., "On the Optimal Ordering Level of the Lower Echelon in Two-Echelon Inventory Systems," Technical Memorandum Serial TM–80237. Institute for Management Science and Engineering, The George Washington University (1972).

[6] Zacks, S., "On the Optimality of the Bayes Prediction Policy in Two-Echelon Multistation Inventory Models," Nav. Res. Log. Quart. *21*, 569–574 (1974).

[7] Zacks, S. and J. Fennell, "Bayes Adaptive Control of Two-Echelon Inventory Systems, I: Development for a Special Case of a One-Station Lower Echelon and Monte Carlo Simulation," Nav. Res. Log. Quart. *19*, 15–28 (1972).

POSTOPTIMALITY ANALYSIS IN INTEGER PROGRAMMING BY IMPLICIT ENUMERATION: THE MIXED INTEGER CASE

Gary M. Roodman

The Amos Tuck School of Business Administration
Dartmouth College

ABSTRACT

This paper develops a method for doing postoptimality analysis on the mixed integer programming problem. The proposed procedures form a natural adjunct to enumerative I.P. algorithms that are linear programming based, and they are designed, in effect, to capitalize on insights generated as the problem is initially solved to do subsequent analysis upon it. In particular, limited ranging analysis is possible on selected parameters, as is the efficient resolving of the problem following parameter changes.

I. INTRODUCTION

Postoptimality analysis for the pure zero-one programming problem has been discussed in Reference [14] using an extension of Balas' Additive Algorithm. The procedures developed there are designed, in effect, to capitalize on insights generated as the problem is initially solved to do subsequent analysis upon it. The procedures permit limited ranging analysis on each parameter and efficient resolving of the problem following parameter changes. Related work in this area has been done by Gomory and Baumol [8], Jensen [9], Noltemeier [12], and Piper and Zoltners [13].

The purpose of this paper is to develop methods of postoptimality analysis for the *mixed integer* programming problem, using in this case an enumerative algorithm that is linear programming based. Such enumerative (or "branch and bound") methods are represented in the work of Land and Doig [10], Dakin [2], Davis, Kendrick, and Weitzman [4], Tomlin [15], and others. Section II of the paper reviews some concepts that are essential in all enumerative algorithms that are linear programming based and states the principal "Fathoming" tests for such algorithms. Section III then outlines an approach to postoptimality analysis for the mixed integer problem which is similar in philosophy, although in few of its specifics, to the work reported in [14]. In section IV, an algorithm embodying the approach is stated, and certain improvements in the method are outlined in section V. Section VI is devoted to an examination of the special problems entailed in using penalty functions, a device that can greatly strengthen the basic algorithm. The paper concludes with a brief discussion of the implementation of the procedures on the time-sharing computer.

II. ENUMERATIVE METHODS

Consider the mixed integer programming problem of the form

(1) Maximize: $z = \mathbf{c}'\mathbf{x}$

P:

$$\text{Subject to: } A\mathbf{x} \leq \mathbf{b} \tag{2}$$

$$x_j = 0 \text{ or } 1, \quad j = 1, \ldots, p \tag{3}$$

$$0 \leq x_j \leq 1 \quad j = p+1, \ldots, n \tag{4}$$

where $A = ||a_{ij}||$ is an $m \times n$ matrix, and $\mathbf{c} = (c_1, \ldots, c_n)$, $\mathbf{x} = (x_1, \ldots, x_n)$, and $\mathbf{b} = (b_1, \ldots, b_m)$ are column vectors. For simplicity the problem has been stated here with the x_j all restricted to the [0, 1] interval, but the results of the paper can be extended to the more general case in which "x_j integer, $j = 1, \ldots, p$" replaces (3) and "$x_j \geq 0, j = p+1, \ldots, n$" replaces (4).

We shall define a *partial assignment*, S, as a specification of (0, 1)-values for a subset of the p binary variables in P. Binary variables that are in the partial assignment are termed *fixed*, while all other binary variables are *free*. When x_j is in S, its value (0 or 1) is denoted by x_j^S. It will be convenient to define the two index sets

$$J^S = \{j | x_j \text{ fixed with respect to } S\}$$

and

$$\bar{J}^S = I_p - J^S, \text{ where } I_p = \{1, \ldots, p\}.$$

A *completion* of S is then defined as any $\mathbf{x}$ for which $x_j = x_j^S$, $j \in J^S$. Note that, in general, most completions of S will not be feasible solutions of P (i.e., they will not satisfy (2), (3), and (4)).

Enumerative methods of integer programming constitute one approach to solving the mixed integer-continuous problem (see Geoffrion and Marsten [7]). All enumerative algorithms for solving P involve generating a sequence of partial assignments and studying the completions of each in search of successively better feasible solutions. There are many systematic ways to choose a sequence of partial assignments that affects a nonredundant, but exhaustive search of all solutions to P (see e.g., Balas [1], Geoffrion [6], Land and Doig [10], and Little, Murty, Sweeney, and Karel [11]).

For enumerative algorithms that are linear programming based, two optimization problems are associated with each partial assignment, S. The first is the *candidate problem*, P^S, defined for S:†

$$\text{Maximize: } z = \sum_{j \in \bar{J}^S} c_j x_j + \sum_{j=p+1}^{n} c_j x_j + z^S.$$

P^S:

$$\text{Subject to: } \sum_{j \in \bar{J}^S} a_{ij} x_j + \sum_{j=p+1}^{n} a_{ij} x_j \leq b_i^S$$

$$x_j = 0 \text{ or } 1, \quad j \in \bar{J}^S$$

$$0 \leq x_j \leq 1 \quad p+1, \ldots, n,$$

†The terminology and much of the notation in this paper is borrowed from Geoffrion and Marsten [7].

where $z^S = \sum_{j \in J^S} c_j x_j^S$ and $b_i^S = b_i - \sum_{j \in J^S} a_{ij} x_j^S$ are constants. Problem P^S is simply the mixed integer programming problem that remains after the variables in S have been assigned their fixed values. It is clear that every (in)feasible solution of P^S, together with S itself, yields a (in)feasible solution of P.

The second problem of interest is the *relaxed candidate problem*, P_r^S, formed from P^S by relaxing the requirement "$x_j = 0$ or 1, $j \in \bar{J}^S$" to "$0 \leq x_j \leq 1$, $j \in \bar{J}^S$". Problem P_r^S is simply a linear programming problem. Its usefulness lies in the inferences that can often be drawn about the solutions of P^S (and hence about the solutions of P itself) from an examination of P_r^S. The inferences all depend upon the fact that the feasible solution space of P^S is wholly contained in that of P_r^S.

Let $\bar{z}$ be the value of the best feasible solution of P that is known at the point when S and its completions are under examination. This solution, hereafter called the *incumbent*, may or may not be an optimal solution of P. In addition, let $v(P^S)$ and $v(P_r^S)$ be the values of the optimal solutions of P^S and P_r^S, respectively. Then one of the following conditions may be observed in solving P_r^S:

(a) $v(P_r^S) \leq \bar{z}$, in which case no completion of S yields a feasible solution of P that is better than the incumbent (since $v(P^S) \leq v(P_r^S)$).

(b) *The optimal solution for* P_r^S *is feasible in* P^S (in particular, the x_j, $j \in \bar{J}_S$, assume integer values). This solution must be the best feasible solution that can be generated as a completion of S.

(c) P_r^S *has no feasible solution,* in which case S has no completion that is feasible in P.

Conditions a, b, and c are central to all enumerative algorithms that are linear programming based. When a partial assignment is found that satisfies one of these conditions, it is said to have been *fathomed* and its completions are said to have been *implicitly enumerated.*

III. POSTOPTIMALITY ANALYSIS

We wish now to demonstrate that the use of linear programs in the tests just outlined makes it possible, each time a partial assignment is fathomed, to obtain insights into the structure of the mixed integer problem that are useful in doing postoptimality analysis. We shall deal here only with the analysis of the contribution coefficients and right-hand sides, although the arguments can be extended to the constraint coefficients as well.

The Approach

Consider the problem T, constructed from P:

T:

$$\text{Maximize: } z = \mathbf{c}^{+\prime}\mathbf{x}$$

$$\text{Subject to: } A\mathbf{x} \leq \mathbf{b}^{+}$$

$$x_j = 0 \text{ or } 1, \qquad j = 1, \ldots, p$$

$$0 \leq x_j \leq 1, \qquad j = p+1, \ldots, n$$

where c_j^+ is the smallest value of c_j of interest, and b_i^+ is the smallest b_i of interest. Problem T is the tightest (most conservative) formulation of P to be studied. Parameter changes, if any are made, will involve increasing right-hand sides or increasing contribution coefficients, implying potentially *better* optimal solutions. The analogues of P^S and P_r^S for T may be written as T^S and T_r^S, respectively.

Let Λ denote the set of partial assignments fathomed in solving T by an enumerative procedure that uses the three tests outlined in section II, and let κ be the number of elements in Λ. As long as the search scheme employed is exhaustive, but nonredundant, every solution of T, feasible or infeasible, will be a completion of precisely one element of Λ, so that Λ, in effect, partitions the solutions of T into κ disjoint sets. Each of these sets was studied and then put aside (implicitly enumerated) in the process of solving T.

Once Λ has been defined, it is possible, whenever a parameter of T is changed, to execute an exhaustive, nonredundant search for the new optimum by simply reexamining every element of Λ. For each $S \epsilon \Lambda$, this would require reconstructing T_r^S using the new parameter value and then resuming the study of completions of S in search of feasible solutions *better* than the initial optimum. While this process could be tantamount, in computational terms, to resolving the entire problem, it will generally not be—for two reasons. First, as a general matter, one may reasonably expect the partitioning that yielded the optimal solution to T to provide a valuable perspective for viewing the revised problem as well (it was from this viewpoint that the original problem was "divided and conquered"). Second and more important, however, is the fact that, with a bit of additional analysis as each partial assignment is initially fathomed, a basis can be established for disregarding in each reoptimization all but a small subset of the elements of Λ. In particular, when partial assignment S is initially fathomed, information can be accumulated with which to establish a lower bound on the change in each parameter that would be just sufficient to invalidate the fathoming condition. Only when a parameter change exceeds the appropriate bound would S require full reexamination. Such bounds thus permit efficient solution of the modified problem. In addition, they provide the means for doing ranging analysis on the parameters of the original problem.

Notation

We shall utilize the following notation throughout:

1. Let $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ be the *initial* optimal solution of T, and let z^* be the associated objective function value (z^* is the value of the final incumbent).

2. Let $z^*(\Delta \cdot)$ be the objective function value for the *initial* optimum, $\mathbf{x}^*$, following parameter change $\Delta \cdot$. Since no admissible parameter change will render $\mathbf{x}^*$ infeasible, we may take †

$$z^*(\Delta c_j) = z^* + x_j^* \cdot \Delta c_j \tag{5}$$

$$z^*(\Delta b_i) = z^*. \tag{6}$$

3. Let $T(\Delta \cdot)$ denote problem T reformulated to incorporate parameter change $\Delta \cdot$, and let $T^S(\Delta \cdot)$ and $T_r^S(\Delta \cdot)$ be the associated candidate and relaxed candidate problems, respectively, for $S \epsilon \Lambda$.

4. Let Λ_a, Λ_b, Λ_c be the set of partial assignments fathomed by conditions a, b, and c, respectively, in solving T ($\Lambda = \Lambda_a \cup \Lambda_b \cup \Lambda_c$).

† Alternatives to $z^*(\Delta \cdot)$ are discussed in section V.

5. For each $S\epsilon\Lambda_a\cup\Lambda_b$, let †

   $\mathbf{y}^S=(y_1^S,\ .\ .\ .,\ y_n^S)$ be the optimal solution of T_r^S $(y_j^S=x_j^S,\ j\epsilon J^S)$,

   δ_j^S be the largest increase in

   c_j for which $\mathbf{y}^S$ remains an optimal solution of T_r^S (for $j\epsilon J^S$, we may take δ_j^S equal to an arbitrarily large number),

   γ_i^S be the largest increase in b_i for which the basis yielding $\mathbf{y}^S$ remains feasible in T_r^S, and

   π_i^S be the dual evaluator of constraint i in T_r^S at $\mathbf{y}^S$.

6. For each $S\epsilon\Lambda_c$, let ††

   θ_i^S be the largest increase in b_i for which the (infeasible) terminating solution of T_r^S remains the terminating solution (no further iterations can be undertaken).

7. Finally, for each $S\epsilon\Lambda$, let

   ψ_j^S be a critical value of Δc_j, such that for $\Delta c_j\leqslant\psi_j^S$, no solution of T^S (Δc_j) will yield a feasible solution to $T(\Delta c_j)$ better than $\mathbf{x}^*$.

   ϕ_i^S be a critical value of Δb_i, such that for $\Delta b_i\leqslant\phi_i^S$, no solution of $T^S(\Delta b_i)$ will yield a feasible solution to $T(\Delta b_i)$ better than $\mathbf{x}^*$.

A few additional symbols will be defined as required.

Fathoming by Condition *a*

For $S\epsilon\Lambda_a$, $v(T_r^S)\leqslant z^*$. In locating an optimal solution for $T(\Delta\cdot)$, $T^S(\Delta\cdot)$ will require examination only if $\Delta\cdot$ is sufficiently large to make $v(T_r^S(\Delta\cdot))>z^*(\Delta\cdot)$. We wish to estimate, for each parameter, the minimum $\Delta\cdot$ for which this will occur.

Precisely determining this minimum $\Delta\cdot$ for a given parameter in T_r^S will, in general, require repeatedly incrementing the parameter and reoptimizing T_r^S (using L.P. ranging analysis and restarting techniques), until the $\Delta\cdot$ such that $v(T_r^S(\Delta\cdot))=z^*(\Delta\cdot)$ is located. Since these calculations could be carried out only *after* $\mathbf{x}^*$ and z^* are known, provision would need to be made for "recapturing" each T_r^S, $S\epsilon\Lambda_a$, after problem T has been solved. While the bookkeeping and analysis required here is perfectly straightforward, it would clearly be computationally costly to undertake for all $n+m$ parameters in every T_r^S, $S\epsilon\Lambda_a$.

As an alternative, a *lower bound* on the minimum $\Delta\cdot$ for each parameter can be obtained easily (and almost without cost) by using just the optimal T_r^S solution and the postoptimality analysis that can be done with respect to it. It is necessary simply to note and save certain characteristics of T_r^S as S is initially fathomed, and then combine these characteristics with $\mathbf{x}^*$ and z^*, when they are known, to derive the desired bounds. †††

---

† The δ_j^S and γ_i^S can be generated by the standard methods of linear programming ranging analysis.

†† The determination of θ_i^S will depend on whether the simplex or dual simplex method is being employed. In the simplex method, infeasibility is detected when artificial variables are found with positive values in the "optimal" solution. in this case, θ_i^S will assume the largest value of Δb_i for which that final solution remains feasible. In the dual simplex method. infeasibility is detected when a basic variable having a negative value cannot be forced to zero. In this case, θ_i^S will assume the value of Δb_i for which the negative value goes to zero.

†††A second lower bound on the minimum $\Delta\cdot$ might be generated using the iterative procedure initially outlined, but substituting for the optimal solution the incumbent when the partial assignment is fathomed (thus making the "recapturing" unnecessary). This possibility was tested and found (not surprisingly) to involve an enormous computational burden, with no clear improvement in the bounds generated.

To establish these *lower* bounds, it is necessary to establish *upper* bounds on $v(T_r^S(\Delta c_j))$ and $v(T_r^S(\Delta b_i))$. This can be accomplished using two elementary results from the theory of linear programming (Dantzig [3]):

1. $v(T_r^S(\Delta c_j))$ is a piece-wise linear, *convex* function of Δc_j for every j, $j=1, \ldots, n$.
2. $v(T_r^S(\Delta b_i))$ is a piece-wise linear, *concave* function of Δb_i for every i, $i=1, \ldots, m$.

Thus, for partial assignment S, $v(T_r^S(\Delta c_j))$ and $v(T_r^S(\Delta b_i))$ might appear as the solid lines in Figure 1.

For $0 \leq \Delta c_j \leq \delta_j^S$, we know that

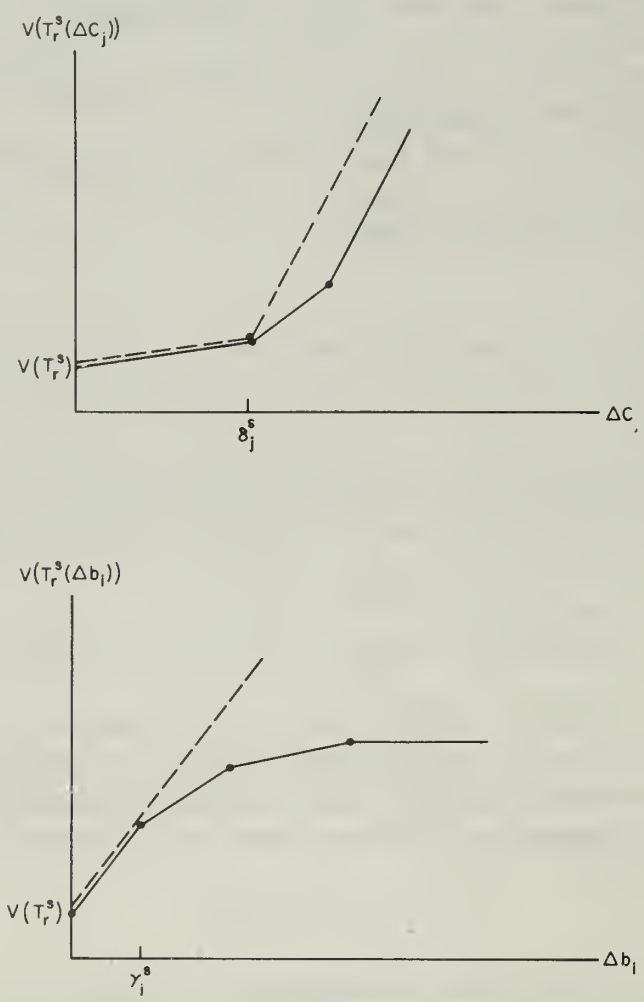


FIGURE 1. (a). Upper bound for $v(T_r^S(\Delta c_j))$. (b). Upper bound for $v(T_r^S(\Delta b_i))$.

$$v(T_r^S(\Delta c_j)) = v(T_r^S) + y_j^S \cdot \Delta c_j. \tag{7a}$$

Furthermore, by definition, $x_j \leq 1$ in any feasible solution of $T_r^S(\Delta c_j)$. Hence, for $\Delta c_j > \delta_j^S$, it will always be true that

(7b) $$v(T_r^S(\Delta c_j)) \leq [v(T_r^S) + y_j^S \cdot \delta_j^S] + (\Delta c_j - \delta_j^S).$$

(In effect, the slope of the "steepest piece" of $v(T_r^S(\Delta c_j))$ cannot exceed one.) Taken together, (7a) and (7b) provide an upper bound on $v(T_r^S(\Delta c_j))$ for all Δc_j. Such an upper bound is represented by the dashed lines in Figure 1a.

For $0 \leq \Delta b_i \leq \gamma_i^S$, we know that

$$v(T_r^S(\Delta b_i)) = v(T_r^S) + \pi_i^S \cdot \Delta b_i.$$

The concavity of $v(T_r^S(\Delta b_i))$ implies, however, that

(8) $$v(T_r^S(\Delta b_i)) \leq v(T_r^S) + \pi_i^S \cdot \Delta b_i$$

for *all* Δb_i, so that (8) provides the required upper bound on $v(T_r^S(\Delta b_i))$. This bound is represented by the dashed line in Figure 1b.

Increase in c_j. For every contribution coefficient, c_j, and every $S \epsilon \Lambda_a$, we wish to find a ψ_j^S such that, for $\Delta c_j \leq \psi_j^S$, $v(T_r^S(\Delta c_j)) \leq z^*(\Delta c_j)$. As long as this condition is satisfied, $T^S(\Delta c_j)$ will not yield a feasible solution of $T(\Delta c_j)$ better than $\mathbf{x}^*$.

From (5), (7a), and (7b), it is clear that $v(T_r^S(\Delta c_j)) \leq z^*(\Delta c_j)$, if $0 \leq \Delta c_j \leq \delta_j^S$ and

(9a) $$v(T_r^S) + y_j^S \cdot \Delta c_j \leq z^* + x_j^* \cdot \Delta c_j,$$

or if $\Delta c_j > \delta_j^S$ and

(9b) $$v(T_r^S) - (1 - y_j^S)\delta_j^S + \Delta c_j \leq z^* + x_j^* \cdot \Delta c_j.$$

For (9a), we may define

$$R_{j1}^S = \begin{cases} [z^* - v(T_r^S)]/[y_j^S - x_j^*], & \text{if } y_j^S \neq x_j^* \\ M, & \text{if } y_j^S = x_j^*, \end{cases}$$

where M is an arbitrarily larger number. For (9b), let

$$R_{j2}^S = \begin{cases} [z^* - v(T_r^S) + (1 - y_j^S)\delta_j^S]/[1 - x_j^*], & \text{if } x_j^* < 1. \\ M, & \text{if } x_j^* = 1. \end{cases}$$

In terms of R_{j1}^S and R_{j2}^S, ψ_j^S may then be defined as

$$\psi_j^S = \begin{cases} R_{j1}^S, & \text{if } 0 \leq R_{j1}^S \leq \delta_j^S \\ R_{j2}^S, & \text{otherwise.} \end{cases}$$

Note that $\psi_j^S = M$, whenever $x_j^* = 1$. This is as one would expect, since $\mathbf{x}^*$ will, in that case, always be an optimal solution of $T(\Delta c_j)$ for any admissible Δc_j.

Increase in b_i. For every right-hand side, b_i, and every $S\epsilon\Lambda_a$, we wish to find a ϕ_i^S such that, for $\Delta b_i \leq \phi_i^S$, $v(T_r^S(\Delta b_i)) \leq z^*(\Delta b_i)$. It follows from (6) and (8) that this condition will be satisfied for all Δb_i such that

$$v(T_r^S) + \pi_i^S \cdot \Delta b_i \leq z^*. \tag{10}$$

Since $\pi_i^S \geq 0$, we may simply define ϕ_i^S by

$$\phi_i^S = \begin{cases} [z^* - v(T_r^S)]/\pi_i^S, & \text{if } \pi_i^S > 0 \\ M, & \text{if } \pi_i^S = 0. \end{cases}$$

Fathoming by Condition *b*

For $S\epsilon\Lambda_b$, the optimal T_r^S solution is feasible in T^S and yields a potential new incumbent. Since the incumbent solution is continually changing, with $\bar{z}$ increasing, every partial assignment fathomed in this way will satisfy the condition $v(T_r^S) \leq z^*$, with the equality holding for the final incumbent. Consequently, for purposes of postoptimality analysis, partial assignments fathomed by Condition *b* may be treated exactly like those fathomed by Condition *a*.

Fathoming by Condition *c*

For $S\epsilon\Lambda_c$, T_r^S has no feasible solution. This condition can be reversed only be relaxing one or more constraints; changes in contribution coefficients will have no effect. To determine exactly the minimal Δb_i that would restore feasibility, one could repeatedly increment b_i and reoptimize until feasibility is achieved. As before, however, this process will generally have high computational costs; and also as before, a lower bound on the minimal change is readily available from the infeasible solution that terminates study of T_r^S.

Increase in b_i. Feasibility will not be attained in T_r^S (Δb_i), as long as $\Delta b_i < \theta_i^S$. At $\Delta b_i = \theta_i^S$, the terminating solution for T_r^S may become feasible, so that we must take

$$\phi_i^S = \theta_i^S - \epsilon,$$

where ϵ is an appropriately small, positive constant. If, for example, b_i always assumes an integer value, then any ϵ, $0 < \epsilon \leq 1$, may be chosen.

IV. THE ALGORITHM

It is useful to view the computational process in two phases, a *solution phase*, in which T is solved and the required information is accumulated, and a *postoptimal phase*, in which ranging analysis can be undertaken and parameter changes can be efficiently studied.

SOLUTION PHASE: Formulate and solve problem T using an enumerative algorithm based upon the fathoming tests of section II. Observe one modification of the usual procedure: when partial assignment S is fathomed, save S (constructing Λ_a, Λ_b, and Λ_c) and the following auxiliary information:

(a) $v(T_r^S)$, y_j^S and δ_j^S, $j = 1, \ldots, n$, and π_i^S, $i = 1, \ldots, m$, if $S\epsilon\Lambda_a \cup \Lambda_b$; or

(b) θ_i^S, $i = 1, \ldots, m$, if $S\epsilon\Lambda_c$.

At the conclusion of the solution phase, $\psi_j^S, j=1, \ldots, n$, may be computed for each $S\epsilon\Lambda_a \cup \Lambda_b$ and ϕ_i^S, i 1, . . ., m, may be computed for each $S\epsilon\Lambda_a \cup \Lambda_b \cup \Lambda_c$, using the auxiliary information and the $\mathbf{x}^*$ and z^* that have been determined. These critical values for each $S\epsilon\Lambda$, together with Λ itself, are then sufficient in the second phase to execute the postoptimality analysis; however, to accelerate the study of parameter changes in the postoptimal phase, it is useful, whenever computer storage permits, to carry forward the entire set of auxiliary information for all $S\epsilon\Lambda_a \cup \Lambda_b$ (rather than the ψ_j^S), so that the ψ_j^S may be recomputed as successively better incumbents are found. The postoptimality procedures, as stated below, assume that the full set of information has been carried forward. (The modifications required when this cannot be done will be apparent.)

POSTOPTIMAL PHASE: Ranging Analysis

(a) On c_j: Let

$$d_j = \min_{S\epsilon\Lambda_a\cup\Lambda_b} \{\psi_j^S\}.$$

The current optimal solution will remain optimal for any $c_j \leq c_j^+ + d_j$.

(b) On b_i: Let

$$g_i = \min_{S\epsilon\Lambda_a\cup\Lambda_b\cup\Lambda_c} \{\phi_i^S\}.$$

If s_i is the amount of slack in constraint i at the current optimum, the current optimal solution will remain optimal for b_i in the interval

$$b_i^+ - s_i \leq b_i \leq b_i^+ + g_i.$$

POSTOPTIMAL PHASE: Parameter Changes

(a) *Revision of* c_j: For each $S\epsilon\Lambda_a \cup \Lambda_b$ in which $\Lambda c_j > \psi_j^S$, formulate and solve $T_r^S(\Delta c_j)$. Take as the initial incumbent the best feasible solution previously encountered (in the solution phase or in any partial assignment previously examined). Any $S\epsilon\Lambda_a\cup\Lambda_b$ for which $\Delta c_j \leq \psi_j^S$ may be disregarded. When every $S\epsilon\Lambda_a\cup\Lambda_b$ has been considered, a new optimal solution, if one exists, will have been found.

(b) *Revision of* b_i: For each $S\epsilon\Lambda$ in which $\Delta b_i > \phi_i^S$, formulate and solve $T_r^S(\Delta b_i)$. Take as the initial incumbent the best feasible solution previously encountered. Any $S\epsilon\Lambda$ for which $\Delta b_i \leq \phi_i^S$ may be disregarded. When every $S\epsilon\Lambda$ has been considered, a new optimal solution, if one exists, will have been found.

V. A MODIFIED METHOD

Let T^* denote the "optimal" linear programming problem formed from T by fixing $x_j = x_j^*, j=1, \ldots, p$. The optimal solution of T^* is simply $\mathbf{x}^*$, with $v(T^*) = z^*$. In addition, let $T^*(\Delta\cdot)$ denote problem T^* reformulated to incorporate parameter change $\Delta\cdot$. The optimal solution of $T^*(\Delta\cdot)$, $\mathbf{y}^*(\Delta\cdot)$, has the useful property that it will always be (1) feasible in $T(\Delta\cdot)$ and (2) at least as good a solution for $T(\Delta\cdot)$ as $\mathbf{x}^*$, since $v(T^*(\Delta\cdot)) \geq z^*$ for any admissible $\Delta\cdot$. Hence, $\mathbf{y}^*(\Delta\cdot)$ and $T^*(\Delta\cdot)$ can be used in several ways to strengthen the postoptimality procedures outlined earlier:

1. The quantity $[v(T^*(\Delta\cdot)) - z^*]$ is an estimate of (a lower bound on) the improvement in the objective function associated with parameter change $\Delta\cdot$ (since $v(T(\Delta\cdot)) \geq v(T^*(\Delta\cdot)) \geq z^*$). Hence, in the "neighborhood" of $\mathbf{x}^*$, it can be used much as dual evaluators are used in linear programming.

2. The procedures stated in section IV take $\mathbf{x}^*$ as the initial incumbent in solving $T(\Delta \cdot)$ following parameter change $\Delta \cdot$. It will generally be beneficial to substitute $\mathbf{y}^*(\Delta \cdot)$ for $\mathbf{x}^*$ as the initial incumbent. (Determining $\mathbf{y}^*(\Delta \cdot)$ will ordinarily require only a few simplex method iterations, once $\mathbf{x}^*$ and the associated basic feasible solution of T^* are known.)

3. For small $\Delta \cdot$, $\mathbf{y}^*(\Delta \cdot)$ will, in general, *be* the optimum for $T(\Delta \cdot)$ (i.e., the initial and revised optimum for $T(\Delta \cdot)$ will differ only in the values of their continuous variables). Hence, it may be possible to *anticipate* the optimal solution of $T(\Delta \cdot)$, without formal reoptimization, even when that solution is not precisely $\mathbf{x}^*$. To identify, for each parameter, a range of values for which $\mathbf{y}^*(\Delta \cdot)$ is the optimal solution of $T(\Delta \cdot)$, it is necessary simply to substitute $v(T^*(\Delta \cdot))$ for $z^*(\Delta \cdot)$ in computing the ψ_j^S and ϕ_i^S for each $S \in \Lambda_a \cup \Lambda_b$. Thus, for example, in place of (10) in section III, the relationship

$$v(T_r^S) + \pi_i^S \cdot \Delta b_i \leq v(T^*(\Delta b_i)) \tag{10'}$$

would be examined in defining ϕ_i^S; and the interval defined by g_i would be an interval over which the optimal solution of $T(\Delta b_i)$ would be known and could be perfectly "tracked" using standard methods of linear programming (on problem T^*).

To make full use of $v(T^*(\Delta \cdot))$ in the modified ranging analysis, it will generally be necessary to examine a sequence of optimal solutions of $T^*(\Delta \cdot)$ for each parameter (using L.P. ranging analysis and restarting techniques) in order to trace out the required portion of the $v(T^*(\Delta \cdot))$ function. Depending upon the number of distinct solutions this process involves, the computational burden could be quite large. Whenever this is the case, however, the examination of solutions for $T^*(\Delta \cdot)$ can be restricted to, say, one new solution for each parameter, so that the optimum for $T(\Delta \cdot)$ is simply being "anticipated" over smaller ranges of values.

At a minimum, useful information can be drawn directly from the optimal solution of T^*. If γ_i^* is the largest increase in b_i for which the basis yielding $\mathbf{x}^*$ remains feasible in T^*, then

$$v(T^*(\Delta b_i)) = z^* + \pi_i^* \cdot \Delta b_i \qquad \text{for } 0 \leq \Delta b_i \leq \gamma_i^*,$$

where π_i^* is the dual evaluator of constraint i in T^* at $\mathbf{x}^*$. Let

$$z^{**}(\Delta b_i) = \begin{cases} z^* + \pi_i^* \cdot \Delta b_i & \text{if } 0 \leq \Delta b_i \leq \gamma_i^* \\ z^* + \pi_i^* \cdot \gamma_i^* & \text{if } \Delta b_i > \gamma_i^*. \end{cases}$$

As defined, $z^{**}(\Delta b_i)$ forms a *lower* bound on $v(T^*(\Delta b_i))$, i.e., $v(T^*(\Delta b_i)) \geq z^{**}(\Delta b_i)$, for *all* Δb_i (see Figure 2(b)). Hence, in place of (10'), we may use the relationship

$$v(T_r^S) + \pi_i^S \cdot \Delta b_i \leq z^{**}(\Delta b_i)$$

to define ϕ_i^S. As before, $\mathbf{y}^*(\Delta b_i)$ will be the optimal solution of $T(\Delta b_i)$ for all Δb_i in the resulting ranging interval. Whenever $g_i > \gamma_i^*$, however, there will be a portion of the interval $(\gamma_i^* < \Delta b_i \leq g_i)$ over which $\mathbf{y}^*(\Delta b_i)$ will not have been determined.

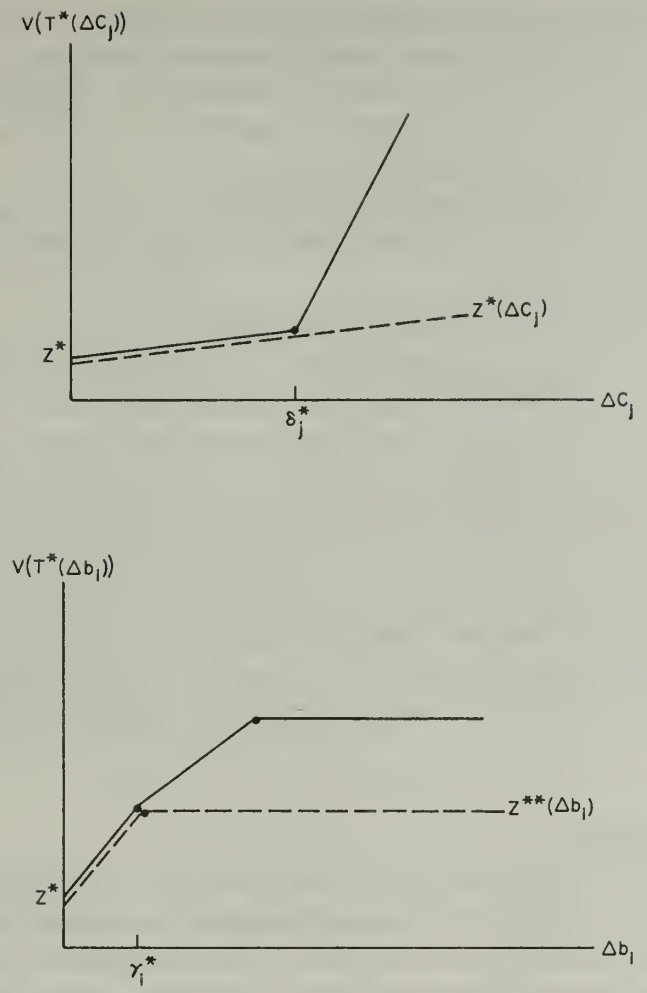


FIGURE 2. (a). Lower bound for $v(T^*(\Delta c_j))$. (b). Lower bound for $v(T^*(\Delta b_i))$.

Using T^* in the ranging analysis on contribution coefficients involves only a very simple extension of the procedures outlined in sections III and IV, since the $z^*(\Delta c_j)$ defined earlier provides a lower bound on $v(T^*(\Delta c_j))$ for *all* Δc_j (see Figure 2(a)). Consequently, the ψ_j^S and d_j can be generated just as before. Over the ranging interval defined by d_j, $\mathbf{y}^*(\Delta c_j)$ will be the optimal solution of $T(\Delta c_j)$. As above, however, when $d_j > \delta_j^*$, where δ_j^* is the largest increase in c_j for which $\mathbf{x}^*$ remains optimal in T^*, there will be a portion of the interval $(\delta_j^* < \Delta c_j \leq d_j)$ over which $\mathbf{y}^*(\Delta c_j)$ will not have been determined.

VI. PENALTY FUNCTIONS

Penalty functions provide one means for strengthening the basic enumerative algorithm employed in earlier sections (see, for example, Davis, Kendrick, and Weitzman [4] and Tomlin [15]). Let S denote some *unfathomed* partial assignment. At the optimum for T_r^S, it is generally possible to associate with each binary variable x_j, $j \epsilon \bar{J}^S$, an *up* penalty, $UP_j^S (\geq 0)$ and a *down* penalty, $DP_j^S (\geq 0)$ which estimate, respectively, the effect on $v(T_r^S)$ of forcing x_j to 1 (imposing the constraint "$x_j = 1$") and forcing x_j to 0 (imposing "$x_j = 0$"). The most common way to estimate each penalty is to determine the change in

$v(T_r^S)$ affected by the first dual simplex iteration following imposition of the associated constraint. Measured in this way, the penalties serve two useful "fathoming" functions:

1. If it should be discovered that UP_j^S (or DP_j^S) is undefined (i.e., the first dual simplex iteration cannot be executed), then there can be no feasible completion of S in which $x_j = 1$ (or $x_j = 0$). Thus, x_j may be fixed at 0 (or 1) in all completions of S.

2. When UP_j^S (or DP_j^S) is defined, it provides a lower bound on the actual change in $v(T_r^S)$ associated with imposing "$x_j = 1$" (or "$x_j = 0$"). Thus, whenever $v(T_r^S) - UP_j^S < \bar{z}$ (or $v(T_r^S) - DP_j^S < \bar{z}$), x_j may be fixed at 0 (or 1) in all completions of S.

In the first case, partial assignment S', formed by augmenting S with $x_j^{S'} = 1$ (or $x_j^{S'} = 0$), is fathomed by Condition c; in the second case, S' is fathomed by a modified form of Condition a.

For the postoptimality procedures described in this paper, it has been found that the most straightforward and computationally desirable way to handle partial assignments fathomed using penalty functions is to temporarily impose the "fathomed" constraint (which can be done very easily in the dual simplex method) and then attempt to execute the first dual simplex iteration. (In the process of computing the penalties the choice of pivot row and pivot column will already have been resolved.) If the one iteration can be executed, it will force the objective function value for $T_r^{S'}$ below $\bar{z}$ and yield a solution that is generally infeasible and "better-than-optimal." For purposes of doing postoptimality analysis, however, S' can be made an element of Λ_a and the infeasible solution can be treated accordingly. † If, on the other hand, the one dual simplex iteration cannot be executed, then S' is simply placed in Λ_c.

VIII. IMPLEMENTATION

The procedures outlined in this paper are now being implemented on the Honeywell 635 computer. The integer programming method to which they are attached is essentially the Dakin algorithm [2] with provision for penalty functions. The procedures are being programmed to make maximum use of time-sharing and the opportunity it provides the user to study his problem on an interactive basis.

Computational experience to date has been very encouraging, although the computer programs do not yet take advantage of some obvious methods for improving efficiency (the Λ file is not sorted, for example, and key simplex tableaus are not saved). For the problems tested to date, there has been a 5- to 10-percent increase in computation time in the solution phase to handle the special analysis and bookkeeping required to save fathomed partial assignments. For small to moderate parameter changes (10 to 50 percent) in the postoptimal phase, however, the time required to locate a new optimum has consistently been a fraction of the time that would have been required if the original problem had been modified and solved again. In one problem, for example, involving 15 binary variables, 13 continuous variables, and 10 constraints, six of the contribution coefficients were each increased by 30 percent. To locate the six new optima, the postoptimality procedures required a total of 9 CPU seconds (57 simplex iterations). Solving the six problems separately, initializing each search with the original optimum, required a total of 37 CPU seconds (219 simplex iterations). For large parameter changes, the

† This statement follows from the fact that the objective function for any dual-feasible solution of $T_r^{S'}$ provides an upper bound on the value of the optimal (feasible) solution of $T_r^{S'}$. The terminating solution here will be dual-feasible for all Δb_i and for Δc_j in the interval over which the optimality conditions remain satisfied.

computational results have been good but, not surprisingly, less dramatic. Even in these cases, however, the postoptimality procedures have never required more time to solve a problem than complete resolution would have involved.

REFERENCES

[1] Balas, E., "An Additive Algorithm for Solving Linear Programs with Zero-One Variables," Operations Research *13*, 517–546 (1965).

[2] Dakin, R. J., "A Tree Search Algorithm for Mixed Integer Programming Problems," Computer Journal *8*, 250–255 (1965).

[3] Dantzig, G. B., *Linear Programming and Extensions* (Princeton University Press, Princeton, New Jersey, 1973).

[4] Davis, R. E., D. A. Kendrick, and M. Weitzman, "A Branch-and-Bound Algorithm For Zero-One Mixed Integer Programming Problems," Operations Research *19*, 1036–1044 (1971).

[5] Driebeek, N. J., "An Algorithm for the Solution of Mixed Integer Programming Problems," Management Science *12*, 576–587 (1966).

[6] Geoffrion, A. M., "Integer Programming by Implicit Enumeration and Balas Method," SIAM Review *9*, 178–190 (1967).

[7] Geoffrion and Marsten, "Integer Programming Algorithms: A Framework and State-of-the-Art Survey," Management Science *18*, 465–491 (1972).

[8] Gomory, R. E., and W. J. Baumol, "Integer Programming and Pricing," Econometrica, *28*, 521–550 (1960).

[9] Jensen, R. E., "Sensitivity Analysis and Integer Linear Programming," The Accounting Review *XLIII*, 425–446 (1968).

[10] Land, A. H., and A. G. Doig, "An Automatic Method of Solving Discrete Programming Problems," Econometrica *28*, 497–520 (1960).

[11] Little, J. D. C., K. G. Murty, D. W. Sweeney, and C. Karel, "An Algorithm for the Travelling Salesman Problem," Operations Research *11*, 972–989 (1963).

[12] Noltemeier, H., "Sensitivitalsanalyse bei disketen linearen Optimierungs-problemen," Lecture Notes in Operations Research and Mathematical Systems, #30, Edited by M. Beckmann, Providence, and H. P. Kunzi, Zurich, Springer-Verlag, Berlin, Heidelberg, New York (1970).

[13] Piper, C. J. and A. A. Zoltners, "Implicit Enumeration Based Algorithms for Postoptimizing Zero-One Programs," Management Science Research Report, No. 313, Carnegie-Mellon University (Mar. 1973).

[14] Roodman, G. M., "Postoptimality Analysis in Zero-One Programming by Implicit Enumeration," Nav. Res. Log. Quart. *19*, 435–447 (1972).

[15] Tomlin, J. A., "Branch and Bound Methods for Integer and Non-Convex Programming," in *Integer and Nonlinear Programming*, edited by J. Abadie (North-Holland, Amsterdam, 1970).

AN EFFICIENT HEURISTIC PROCEDURE FOR THE CAPACITATED WAREHOUSE LOCATION PROBLEM

Basheer M. Khumawala

Graduate School of Business Administration
University of North Carolina at Chapel Hill

ABSTRACT

This paper introduces an efficient heuristic procedure for solving a special class of mixed integer programming problem called the capacitated warehouse (plant) location problem. This procedure parallels the work reported earlier in [9] on the uncapacitated warehouse location problem. The procedure can be viewed as tracing a judiciously selected path of the branch and bound tree (from the initial node to the terminal node) to arrive at a candidate solution. A simple backtracking scheme is also incorporated in the procedure to investigate possible improvement in the solution. Computational results on problems found in the literature look quite encouraging.

1. INTRODUCTION

The simplest form of the capacitated warehouse (plant) location problem has been formulated in the literature [3, 6, 7, 13] as a mixed integer program as follows: with m potential warehouses and n customers,

(P)

$$\text{Minimize } Z = \sum_{ij} C_{ij}X_{ij} + \sum_{i} F_iY_i \tag{1}$$

$$\text{s.t.} \sum_{i} X_{ij} \geq D_j \qquad j = 1, 2 \ldots n \tag{2}$$

$$\sum_{i} X_{ij} \leq S_iY_i \qquad i = 1, 2 \ldots m \tag{3}$$

$$X_{ij} \geq 0 \tag{4}$$

$$Y_i = 0, \quad 1 \text{ (integer)}, \tag{5}$$

where

C_{ij} = the per unit cost of shipping goods from warehouse i to customer j,
D_j = the demand at customer j,
S_i = the capacity of warehouse i,
F_i = the fixed cost associated with warehouse i,
X_{ij} = the decision variable denoting the amount shipped from warehouse i to customer j,
Y_i = the decision variable indicating whether warehouse i is opened ($Y_i = 1$) or closed ($Y_i = 0$).

Note that the thruput from warehouse i, $\sum_j X_{ij}$, can be positive only if potential warehouse i is opened.

When capacity constraints (3) are not present, (i.e., $S_i = \infty$, $\forall_i$) then the problem is called the uncapacitated warehouse location problem. In addition to heuristic approaches [4, 9, 11], exact methods [2, 8, 14, 15] using implicit enumeration techniques have been quite successful (i.e., computationally efficient) in solving the uncapacitated problem. The past research on the capacited problem also consists of heuristic [7, 13] and exact methods [1, 3, 6, 12]. The exact methods are, as in the uncapacitated case, applications of implicit enumeration techniques to the capacitated problem. Davis and Ray [1], Sa [13], and Geoffrion [5] have applied the multibranch and bound technique, whereas, Ellwein [3] developed and applied a modified form of Balas's single branch implicit enumeration method.

It is a well known fact that the computation time and/or computer storage required by the implicit enumeration methods have an exponential relationship to the number of integer variables (potential warehouses). This aspect of the exact methods, therefore, severely limits their use in practice and one resorts to using heuristic approaches. A more mundane reason for using heuristic methods is that in solving location problems of this type, management is rarely interested in the single optimal solution. Management is rather interested in several "good" alternative solutions with an assurance that they are not very far from the optimum. Moreover, as the input data to such problems can never be precise, sensitivity analysis becomes imperative, which requires that the problem be solved repeatedly. It is therefore necessary that a solution method be computationally efficient both in terms of computation time and computer storage.

The solution method presented in this paper possesses these properties. Its computational efficiency enables it to obtain rapidly "good" solutions to large size problems encountered in practice. It is conceptually similar to our earlier procedure [9] for the uncapacitated problem. The mathematical notation and the nomenclature used in the earlier paper are employed here as far as possible.

The remainder of this paper is organized as follows: the next section describes our fundamental solution procedure; section 3 develops the bounds used at each iteration of the solution procedure for fix opening and fix closing of warehouses; section 4 discusses the heuristics derived from these bounds and used in the solution method; section 5 summarizes the solution method and outlines briefly the addition of a simple backtracking scheme; and the final sections give an illustrative problem and the computational results on a variety of test problems taken from the literature.

II. THE FUNDAMENTAL SOLUTION PROCEDURE

In the branch and bound (b&b) technique, a branching decision rule (BDR) is referred to as the

rule by which the "free" integer variable is selected to be constrained. In the warehouse location problem (WLP) the integer variable is Y_i, which represents the status ("open," "closed," or "free" corresponding to $Y_i = 1$, $Y_i = 0$, or $0 \leq Y_i \leq 1$, respectively) of warehouse i. Therefore, in the b&b algorithm for the WLP a free warehouse is selected at a node by a BDR and is constrained both open and closed along the two emanating branches. These branches will be referred to as the open and the closed branches, respectively.

In an earlier paper [8], the author reported several BDR's for the b&b algorithm for the uncapacitated problem. These rules are extended in this research for application to the capacitated problem. These BDR's are of two types, categorized according to whether they favor the opening or closing of the selected free warehouse. Thus the open branch becomes the preferred branch for type I BDR, whereas the closed branch becomes the preferred branch for type II BDR. We introduce the notion of a "preferred path" of a BDR as the path consisting of successive preferred branches starting from the initial node to the terminal node of the b&b tree. ([see 9], p. 114). Therefore, the type I (type II) BDR will have all branches open (closed) along its preferred path. At each node of the preferred path (including the initial node which is common to all preferred paths) certain bounds developed in this study are used by which warehouses can be fixed (as opposed to constrained) open or closed.

Note that the terminal node at the end of any preferred path has all Y_i completely specified zero or one. Hence, the value of the objective function (1) at the terminal node of a path represents a solution (an upper bound) to Problem P. Thus, different BDR's may lead to different solutions to Problem P. The solution method then simply consists of traversing different preferred paths through the application of different BDR's. The solutions are obtained at the end of these preferred paths and the best (minimum) of these solutions (termed the heuristic solution) is obtained.

III. THE BOUNDS

As noted earlier, the bounds to be described in this section are applied at every node of each preferred path for the purpose of (i) fix opening or (ii) closing of free warehouses. This section is accordingly divided into two parts. We shall use the symbols K_0, K_1, and K_2 to denote sets containing the indices of closed, opened, and free warehouses, respectively, at any node. At the initial node, we start with $K_0 = K_1 = \phi$, the empty set and $K_2 = \{1,2, \ldots m\}$. At the terminal node, $K_2 = \phi$ and K_0 and K_1 contain the indices of the closed and open warehouses, respectively.

Bounds for Fix Opening Warehouses

Ellwein [3] applies the following bound (ΔV_i) at the initial node for investigating the possibility of fix opening any free warehouse i: A lower bound on the total variable cost (V Min) that any solution can have is first obtained by solving the transportation problem

$$\text{(T)} \qquad \underset{X_{ij} \geq 0}{\text{Min}} \left\{ \sum_i \sum_j C_{ij} X_{ij} \middle| \sum_i X_{ij} \geq D_j \forall j;\ \sum_j X_{ij} \leq S_i \forall i \right\}.$$

Problem (T) is just problem (P) with all warehouses open for use and their fixed costs assumed zero. Then,

$$\Delta V_i = V\,\text{Min}_i - V\,\text{Min},$$

where $V\,\text{Min}_i$ is the solution to (T) without warehouse i. Clearly, if $\Delta V_i \geq F_i => Y_i = 1$ (warehouse i is also of course opened, if the solution to (T) without warehouse i is infeasible.)*

It is easy to see that the computations of ΔV_i for all i at the initial node require the solutions of a total of $(m+1)$ transportation problems. The procedure which we now discuss avoids solving the transportation problems, but determines powerful lower bounds on ΔV_i, which we shall label $LB\Delta V_i$. Thus, if $LB\Delta V_i \geq F_i => Y_i = 1$. Mathematically, $LB\Delta V_i$ at any node is given by:

for $$i \epsilon K_2$$

let $$\nabla_{ij} = \min_{\substack{k \epsilon K_1 \cup K_2 \\ k \neq i}} [\max (C_{kj} - C_{ij}, 0)],$$

then, $LB\Delta V_i$ is the solution to the following simple knapsack problem:

$$LB\Delta V_i = \text{Min} \Sigma \nabla_{ij} D_j \gamma_j \qquad \text{s.t.} \Sigma D_j \gamma_j \leq S_i \vee_i \qquad 0 \leq \gamma_j \leq 1 \vee_j.$$

The ∇_{ij} have exactly the same meaning as in the uncapacitated case [9], however, the total savings for warehouse i in the capacitated case are limited by S_i, its capacity; and hence the need for solving the knapsack problem.

$LB\Delta V_i$ is clearly much easier to compute than ΔV_i, as it does not require the solution to any transportation problem. The question therefore is how effective is $LB\Delta V_i$ in fix opening warehouses. In a variety of test problems solved in this study, we found that at the initial node its effectiveness was quite resonable when compared against the use of actual bounds ΔV_i. The additional advantage of $LB\Delta V_i$ is the fact that it can be easily applied at any node rather than only at the initial node. †

It is easy to see that the greater the number of free warehouses fixed opened at a node due to $LB\Delta V_i$, the faster will be its convergence to the terminal node (and hence will reduce computation time). In an effort to further increase convergence to the terminal node, we added two *heuristic* "auxiliaries" subsequent to the computation of $LB\Delta V_i$ cost bounds. These auxiliaries have proven quite useful.

Auxiliary No. 1

In computing $LB\Delta V_i$, the capacity S_i of free warehouse i could not be exceeded, even though there may be additional customers which could be best (cheapest) supplied from warehouse ** i, i.e., with $\nabla_{ij} > 0$. If we make a note of such situations in the $LB\Delta V_i$ computations, each free warehouse i would fall in one of the following four categories:

(a) $LB\Delta V_i \geq F_i$ and its capacity is exceeded.
(b) $LB\Delta V_i < F_i$ and its capacity is exceeded.
(c) $LB\Delta V_i \geq F_i$ and its capacity is not exceeded.
(d) $LB\Delta V_i < F_i$ and its capacity is not exceeded.

---

*An obvious fact which Ellwein uses from the information gained from the transportation solution is: Let U_{n+1} denote the optimal dual variable associated with the capacity for warehouse i. Then, clearly, if $U_{n+1} S_i \geq F_i => Y_i = 1$.

†Of course, the actual cost bound ΔV_i can also be made applicable at any node, but as noted earlier, this would require a substantial number of transportation problems to be solved.

** We shall use the term "capacity of warehouse i exceeded" to mean that there was additional demand which could be best assigned to warehouse i, but was not, because of its limited capacity.

The free warehouse i is of course fixed opened in cases (a) and (c). Cases (a) and (b) further indicate that there is some demand which even though is cheapest from warehouse i (in terms of lowest c_{ij}), needs to be assigned to an alternative warehouse. This means that the unassigned demand can now be best supplied from some other warehouses which in turn will increase the $LB\Delta V$ of other warehouses. Thus, the additional step consists of recomputing the $LB\Delta V$ of the remaining free warehouses to see if they now qualify to be fixed opened at this node.

In order to update the $LB\Delta V$ of the remaining free warehouses, the following has to be done first: the warehouses falling in (a) and (c) are fixed opened and their assigned demands are suppressed from further consideration*; also the capacities of warehouses falling in (c) are replaced by their as yet unallocated capacities.

The updated $LB\Delta V_i$ for each of the remaining free warehouses will again fall in one of the four categories (a) through (d) listed earlier. If at least one warehouse is found to fall in category (a), the $LB\Delta V$ updating procedure is cycled again. Thus, the $LB\Delta V$ auxiliary No. 1 routine is executed whenever at least one warehouse is fixed opened whose capacity is exceeded.

Auxiliary No. 2

The second auxiliary for updating $LB\Delta V$ is executed when there is no warehouse falling in category (a), but there is at least one free warehouse belonging to category (b). Note that the category (b) warehouses obviously cannot be fixed opened. But since their "capacity was exceeded," an alternative (second best) warehouse is needed to which such demands can be assigned; and doing so would possibly improve the $LB\Delta V$ of the other warehouses. Thus, the second auxiliary consists of recomputing the $LB\Delta V$ of the remaining free warehouses (excluding the warehouses of category (b), which we shall term "temporarily opened")† to see if they now qualify to be fixed opened at this node.

To accomplish this, the following is done first: the warehouses falling in category (c) are fixed opened, while those in (b) are temproarily opened; their assigned demands are suppressed from further consideration; also the capacities of warehouses falling in (c) are replaced by their as yet unallocated capacities.

The updated $LB\Delta V_i$ for each of the remaining free (excluding the temporarily opened) warehouses will again fall in one of the four categories (a) through (d). If at least one warehouse is found to fall in category (a), the $LB\Delta V$ updating procedure of auxiliary No. 1 is followed after "freeing" the temporarily opened warehouses and releasing their assigned demands. If no warehouse is found to fall in category (a) at the end of the second auxiliary, then the procdure for fix opening of warehouses is ended after "freeing" the temporarily opened warehouses and releasing their assigned demands. The solution procedure then goes to the computations of bounds for possible fix closing of free warehouses, which is discussed next.

It must be noted that these auxiliaries are heuristic and may therefore result in fix opening a warehouse which might in fact be closed in the optimal solution.

---

* Here. by further consideration, we mean only in the computations of $LB\Delta V$ of the free warehouses at this and subsequent nodes.

† These warehouses are made free again and their allocated demand made available as soon as the next updating of the $LB\Delta V$ is completed.

Bound for Fix Closing Warehouses

This bound measures the minimum savings that a free warehouse can generate if it were opened, when considered over the already open warehouses. If these savings do not exceed the fixed cost of the free warehouse, then clearly such a warehouse should be closed. For $i \epsilon K_2$, let

$$\Omega_i' = Z_\tau(K_1) - Z_\tau(K_1 \cup \{i\}),$$

where $Z_\tau(K)$ is the optimal solution (objective function value) to the transportation problem with only the warehouses contained in the set K. Clearly, if $\Omega_i' \leq F_i$ then $Y_i = 0$. Note that, this will never happen if the capacities of the warehouses in set K_1 are not sufficient to supply the total demand.*

We shall use the term "K_1—feasible" to mean the situation when the warehouses in the open set can supply all the customers' demands. In case K_1 is not feasible, the solution $Z_\tau(K_1)$ is obtained by the usual addition of a dummy source with sufficiently large transportation costs. This insures from closing any free warehouse in such case; however, as will be seen later, the information provided by Ω_i' even in such cases is useful in discriminating between free warehouses and has been put to use. Hence, we shall refer to computing Ω_i' regardless of K_1—feasibility.

To compute Ω_i' for all $i \epsilon K_2$ at all nodes requires solutions to a very large number of transportation problems, namely, $Z_\tau(K_1 U\{i\})$ for all $i \epsilon K_2$. Our study again avoids this and suggests to obtain an upper bound on Ω_i' (denote it by $UB\Omega_i$) simply by making use of the already available solution $Z_\tau(K_1)$. Thus, if $UB\Omega_i \leq F_i$, then $Y_i = 0$. $UB\Omega_i$ is obtained as follows:

Let u_r, $r \epsilon K_1$ and V_j $(j = 1, 2 \ . \ . \ . \ n, n+1)$ be the optimum dual variables associated with the solution $Z_\tau(K_1)$ where $n+1$ is the dummy customer. Define,

$$w_{ij} = \text{Min } (C_{ij} + V_{n+1} - V_j, 0), \qquad j = 1, 2 \ . \ . \ . \ n.$$

Note that $w_{ij} \geq 0$ and measures the per unit savings in the variable cost of customer j's demand if supplied from warehouse i. Then, $UB\Omega_i$ is the solution to the following simple knapsack problem:

$$UB\Omega_i = \max \Sigma w_{ij} D_j \delta_j, \qquad \text{s.t. } \Sigma D_j \delta_j \leq S_i \forall i, \qquad 0 \leq \delta_j \leq 1 \ \forall j.$$

As in the case of $LB\Delta V_i$, the $UB\Omega_i$ are also found to be reasonably powerful compared to the exact bound Ω_i' in fix closing warehouses.

This completes the discussion of the bounds computed for both the fix opening and fix closing of the warehouses at any stage. These bounds are applied in a cyclic manner, i.e., whenever a new warehouse is opened (closed), investigation is made if another warehouse can now be closed (opened). The cycle stops when either opening or closing does not occur as a result of the application of the bounds.

IV. HEURISTIC RULES DERIVED FROM THE BOUNDS

Consistent with our earlier paper [9] on the uncapacitated problem and as discussed in sec. II, the various heuristics are derived from the information available from the bounds already computed at the

* Note that infeasibilities may also arise due to the existence of prohibitive routes in the problem.

node. Recall that a BDR is a rule by which the free warehouse to be constrained from the given node is selected. {Each heuristic in turn is associated with the application of the corresponding BDR.} We therefore first discuss the development of these BDR's as they are derived from the bounds $LB\Delta V_i$ and $UB\Omega_i$.

We established that at a node if $LB\Delta V_i \geq F_i$ (or $UB\Omega_i \leq F_i$) for a free warehouse i, then $Y_i = 1$, (or $Y_i = 0$); however, if $LB\Delta V_i < F_i$ (or $UB\Omega_i < F_i$) then warehouse i remains free. Still, in this latter case the information on $LB\Delta V_i$ and $UB\Omega_i$ can be put to "good" use. For convenience, let

$$\Delta_i = LB\Delta V_i - F_i \qquad \text{and } \Omega_i = UB\Omega_i - F_i.$$

Thus, the remaining free warehouses at any node will always have negative Δ_i and positive Ω_i. Intuitively, the larger (smaller) the Δ_i, the greater is the likelihood of warehouse i to be open (closed) at the terminal node reached from the current node. A similar argument can also be made with Ω_i. The following BDR's can therefore be derived from these bounds.*

1. Open the warehouse with the *Largest Delta*
2. Open the warehouse with the *Largest Omega*
3. Close the warehouse with the *Smallest Delta*
4. Close the warehouse with the *Smallest Omega*

Additional branching decision rules can also be formulated. For example, a set of two simple rules can be based on the capacity of the free warehouse. The larger the capacity of the free warehouse, the greater is its possibility of being open and vice versa. Other nonsimple branching decision rules can be obtained, but would require extensive additional computations. e.g.: if at each node the solution $Z_T(K_1 U K_2)$ is also obtained which would yield the slack capacity of warehouse i, then Y_i can be defined as

$$Y_i = 1 - \frac{\text{Slack capacity of warehouse } i.}{S_i}$$

Two branching decision rules can then be based on the Y values. The closer the value of Y to one (zero) the greater is its likelihood to be open (closed) at the terminal node.

One of the objectives of our study was to test the efficiency (in terms of solution quality and computation time) of different heuristic rules. As such, we formulated and tested six heuristic rules which were derived from the following six BDR's.†

1. Open Largest Δ
2. Open Largest Ω
3. Open Largest S
4. Close Smallest Δ
5. Close Smallest Ω
6. Close Smallest S

The six heuristics thus involve the application of the six BDR's and can be represented as traversing their preferred paths. Note that the initial node remains common to all the heuristics. The terminal solutions at the end of the preferred paths become candidates for the heuristic solution to problem (P).

---

*The names of these rules have been carried from our earlier paper [9] on the uncapacitated problem.

† We avoided the use of the Y rules as they require solutions to additional transportation problems and necessitate extra computational effort.

V. PROCEDURE FOR TRAVERSING A PREFERRED PATH

This section now summarizes the solution procedure discussed thus far. We defined a preferred path as a path which starts from the initial node of the b&b tree and connects only the preferred branches (corresponding to its BDR) and ends at a terminal node of the tree. The initial node is characterized as the node at which all potential warehouses are free, whereas the terminal node is one at which there are no free warehouses. If the BDR used is of type I (II), the successive preferred branches of the preferred path will be open (closed) branches. Regardless of the type of the BDR the two bounds are applied at every node of the preferred path (including the initial node) to investigate if any of the free warehouses can either be fixed open or fixed closed. The solution $Z_T(K_1)$ is obtained at the terminal node of each preferred path, and $Z_T(K_1) + \sum_{i \in K_1} F_i$ becomes a candidate solution to problem (P). The application of several BDR's may lead to different candidate solutions. The minimum of these can then be taken as the heuristic solution to problem (P).

A very simple backtracking scheme is also added at the end of each of the preferred paths to investigate any possible improvement in the terminal solution reached. This scheme simply entails reversing the status of the constrained warehouse (i.e., opening the warehouse which is constrained closed and vice versa) one at a time and starting with the last warehouse constrained. In the type I preferred path, this procedure is followed from that node which was K_1-feasible first. This is because all its subsequent nodes down to the terminal node are obviously created only if they improve the incumbent solution.

This backtracking scheme does add to the total computation time for solving the problem, but was found to improve the terminal solution in many problems tested, leading in fact to the optimal solutions in several cases.

VI. ILLUSTRATIVE PROBLEM

Table 1 gives a 5×8 problem that we shall use for illustrating our solution procedure. We shall solve this problem by using the largest omega heuristic only. The solution steps are given below and the preferred path is shown in Figure 1.

TABLE 1. *Illustrative problem*

Potential warehouses	Fixed costs (F_i)	Per unit variable costs = C_{ij} customers								Warehouse capacity
		1	2	3	4	5	6	7	8	
A	100	12	18	10	L[a]	6	L	18	L	25
B	70	21	L	15	24	5.5	21	11	16.5	20
C	60	18	19	11	19.5	5	L	L	19.5	20
D	110	21	19	15	18	6.5	12	16	12	50
E	80	17	15	11	15	7	19.5	20	L	35
Customer demand		15	10	10	15	5	15	20	10	

[a] L indicates a prohibitive route.

1. Initialization

   Initial node number = 0

$$K_0^{(0)} = K_1^{(0)} = \phi, \text{ the empty set, and } K_2^{(0)} = \{A, B, C, D, E\}$$

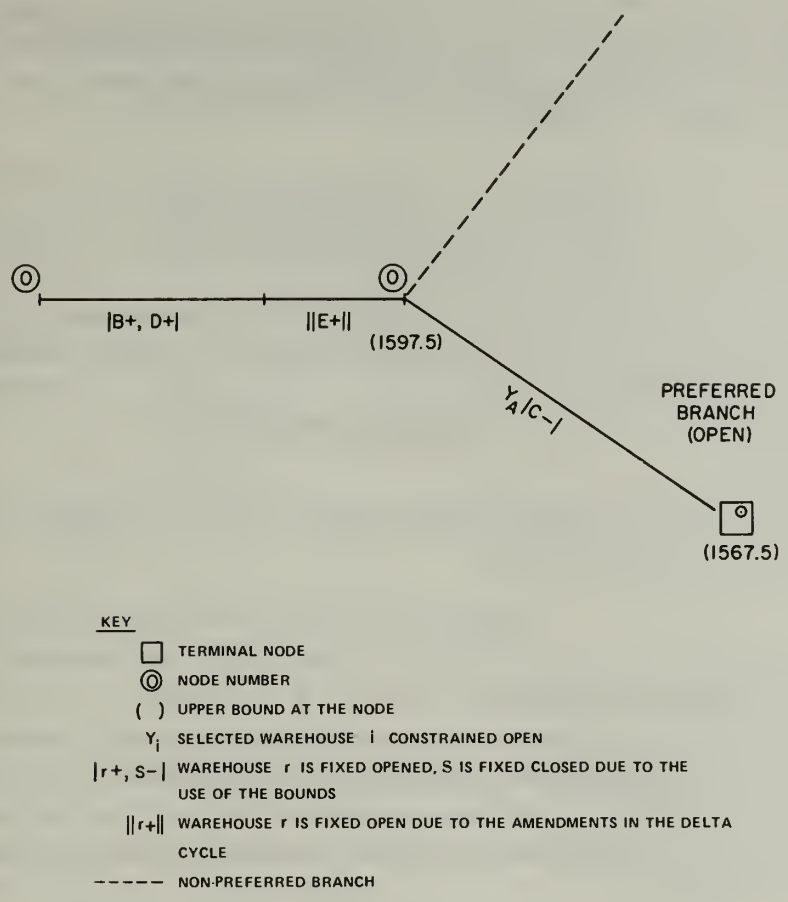


FIGURE 1.

2. Delta Bound Cycle

At this initial stage, the nonzero ∇_{ij}'s are:

$$\begin{aligned}
&\nabla_{A1} = 5, \quad \nabla_{A3} = 1 && \text{yielding } LB\Delta V_A = 85\\
&\nabla_{B7} = 5 && \text{yielding } LB\Delta V_B = 100\\
&\nabla_{C5} = 0.5 && \text{yielding } LB\Delta V_C = 2.5\\
&\nabla_{D6} = 7.5, \ \nabla_{D8} = 4.5 && \text{yielding } LB\Delta V_D = 157.5\\
&\nabla_{E2} = 3, \quad \nabla_{E4} = 3 && \text{yielding } LB\Delta V_E = 75.
\end{aligned}$$

Since the $LB\Delta V$ of warehouses B and D exceed their corresponding fixed costs, warehouse B and D are opened. Furthermore, customers 6, 7, and 8 are eliminated from further consideration at this stage. Also, the fact that the capacity of warehouse B is all exhausted for customer 7 is to be reflected in the subsequent computations of ∇_{ij}'s of the remaining free warehouses.

We now have $K_0 = \phi$, $K_1 = \{B, D\}$ and $K_2 = \{A, C, E\}$ and the new ∇_{ij}'s are:

$$\begin{aligned}
&\nabla_{A1} = 5, \ \nabla_{A3} = 1 && \text{yielding } LB\Delta V_A = 85\\
&\nabla_{C5} = 1 && \text{yielding } LB\Delta V_C = 5\\
&\nabla_{E2} = 3, \ \nabla_{E4} = 3 && \text{yielding } LB\Delta V_D = 75.
\end{aligned}$$

Clearly, no warehouses can now be fixed opened; however, we note that the capacity of warehouse A has been exhausted in the $LB\Delta V$ computations. We therefore apply the auxiliary steps and accordingly temporarily open A and temporarily suppress the total demands of customers 1 and 3. Computing the $LB\Delta V$ now of free warehouses C and E gives:

$$\nabla_{C5} = 1.5 \qquad \text{yielding } LB\Delta V_C = 7.5$$
$$\nabla_{E2} = 4, \quad \nabla_{E4} = 3 \text{ yielding } LB\Delta V_E = 85.$$

Warehouse E is now fixed opened (because $LB\Delta V_E > F_E$) and warehouse A is re-freed and customers 1 and 3 are released. The delta bound cycle now stops, as the capacity of warehouse E was not exceeded in computing its $LB\Delta V$.

At this stage, we now have, $K_0 = \phi$, $K_1 = \{B, D, E\}$ and $K_2 = \{A, c\}$.

3. Omega Bound

We solve the transportation problem with only warehouses B, D, and E, which yields $Z_T(\text{BDE}) = 1337.5$. Since this node is "K_1-feasible" i.e.,

$$S_B + S_D + S_E > \sum_{j=1} D_j,$$

we have an upper bound to our problem, i.e.,

$$UB(\text{node } 0) = Z_T(BDE) + F_B + F_D + F_E = 1597.5.$$

We now use the dual variables from the solution $Z_T(K_1)$ and find the $UB\Omega$ for the free warehouses A and C. The nonzero w_{ij}'s are:

$$w_{A1} = 9, \; w_{A2} = 1, \; w_{A3} = 5, \quad w_{A5} = 0.5, \text{ yielding } UB\Omega_A = 185$$
$$w_{C1} = 3, \; w_{C3} = 4, \; w_{C5} = 1.5 \qquad \text{yielding } UB\Omega_C = 70.$$

Since the $UB\Omega$ of warehouses A and C are not less than the corresponding fixed warehouse costs, these warehouses cannot be closed at this stage.

4. Selection of the Free Warehouse to be Constrained

According to the "largest omega" branching decision rule, we select warehouse A and constrain it to be open. Therefore at node 1 we have

$$K_0 = \phi, \; K_1 = \{A, B, D, E\} \text{ and } K_2 = \{C\}.$$

5. Omega Bound

The transportation solution is $Z_T(ABDE) = 1207.5$, yielding an improved upper bound

$$UB(\text{node } 1) = 1207.5 + 360 = 1567.5.$$

We now use the dual variables in the transportation solution $Z_T(ABDE)$ to obtain $UB\Omega_C$;

$$w_{C5} = 1.5 \text{ yielding } UB\Omega_C = 7.5$$

Thus warehouse C is fix closed and we now have,

$$K_0 = \{C\},\ K_1 = \{A, B, D, E\} \text{ and } K_2 = \phi.$$

The terminal solution is then \$1,567.5

No backtracking can be done, since in this case, the initial node itself was "K_1-feasible." We have therefore reached the Largest Omega Heuristic Solution, which in its completeness is given by:

Open warehouses \ Customers	1	2	3	4	5	6	7	8
A	15	—	10	—	—	—	—	—
B	—	—	—	—	—	—	20	—
D	—	—	—	—	5	15	—	10
E	—	10	—	15	—	—	—	—

VII. COMPUTATIONAL RESULTS

Our solution procedure was programmed in FORTRAN IV and run on IBM 370/165 using the *G*-compiler. The Kuehn and Hamburger (K & H) [11] test problems have now been commonly used by several reserachers. Although these problems were originally used for testing an uncapacitated algorithm, Sa [13] extended their use to test capacitated algorithms as well. Ellwein [3] also used Sa's extended problems and a set of two additional problems derived from an approximate airport location problem. In this study, we employed all of these and a few other problems which are all described in Table 2.

TABLE 2. *Description of test problems*

Problem set	Number of problems	Description	M	N	S_i	F_i
I [a]	4	K & H; Plant: Indianapolis	25	50	58,268	75,000/12,500/17,500/25,000
II	4	K & H; Plant: Jacksonville	25	50	58,268	75,000/12,500/17,500/25,000
III	4	K & H; Plant: Baltimore and Indianapolis	26	50	58,268	75,000/12,500/17,500/25,000
IV [b]	4	K & H; Plant: Indianapolis	24	50	58,268	75,000/12,500/17,500/25,000
V	4	K & H; Plant: Indianapolis	16	50	5,000	75,000/12,500/17,500/25,000
VI	4	K & H; Plant: Indianapolis	16	50	15,000	75,000/12,500/17,500/25,000
VII	4	K & H; Plant: Indianapolis	25	50	5,000	75,000/12,500/17,500/25,000
VIII	4	K & H; Plant: Indianapolis	25	50	15,000	75,000/12,500/17,500/25,000
IX [c]	2	K & H; Plant: Indianapolis	16	50	5,000	12,500/17,500
X [d]	1	Ellwein's [3] problem #2	15	45	1,000–5,000	15,000–40,000
XI [d]	1	Ellwein's [3] problem #3	15	45	1,500–7,500	22,500–60,000

K & H: Kuehn and Hamburger [11]

a: We shall use the convenient notation "problem I–I" to mean the first problem of set I, i.e. with fixed warehouse costs \$7,500.

b: Same as problem set I, except that Indianapolis is not permitted to serve as a warehouse

c: Same as problems V–2,3, except that the Indianapolis Warehouse is also charged with the fixed warehouse cost.

d: The two numbers given under the S_i and F_i columns in these rows indicate the ranges of the corresponding values. These are given in Ellwein [3, p. 88]

Problem sets I through III are the original 12 problems of K & H and were used by Sa as capacitated problems. Problem sets IV, V, VI (1–3) and VII–2 were originally derived by Sa from the K & H problems. Problems IX–2, X and XI are from Ellwein. The remaining problems were our additions, which as can be seen, are extensions of the previously reported problems. We solved all of these sets of problems with our program.

The solutions to problem sets I–IV duplicate * (as they should) those given in our earlier paper [9]; consequently these are not shown here. All the six individual heuristic solutions for each of the remaining problems are given in Table 3. The CPU time (in *seconds*) taken to reach these solutions are also pro-

TABLE 3. *Computational results*

Problem set	Rule No.	Z	t	Z	t	Z	t	Z	t
V (1–4).....	1	1043.3	2.42	1097.9*	1.05	1152.9*	1.07	1244.9	2.69
	2	1043.3	2.39					1235.4*	4.24
	3	1043.3	2.49					1235.4*	10.28
	4	1043.3	2.48					1244.9	2.37
	5	1043.3	3.55					1235.4*	1.24
	6	1044.2	3.60					1292.8	8.40
VI (1–4).....	1	932.5*	0.23	981.4	0.28	1017.3	1.45	1053.1	1.32
	2	932.5*	0.22	977.7*	0.21	1013.9*	1.72	1053.1	1.71
	3	932.5*	0.22	977.7*	0.21	1014.0	1.79	1060.6	1.61
	4	932.5*	0.26	977.7*	0.21	1013.9*	1.61	1058.9	1.05
	5	932.5*	0.22	977.7*	0.24	1017.6	2.17	1077.2	2.29
	6	934.1	0.30	983.6	0.23	1015.5	0.98	1045.5*	0.90
VII (1–4).....	1	842.1	8.37	914.4	12.50	986.2	17.05	1077.9	17.64
	2	848.1	10.72	921.0	15.03	984.0	25.71	1074.0	20.90
	3	846.9	8.49	920.0	12.58	992.4	20.46	1097.4	26.31
	4	838.2*	4.72	912.8	6.06	996.5	6.34	1109.0	9.19
	5	840.6	4.99	921.5	7.64	1011.8	11.81	1078.3	8.82
	6	847.4	8.40	928.3	6.99	1026.5	13.49	1152.8	12.51
VIII (1–4).....	1	796.4*	0.50	860.7	2.85	904.2	4.45	956.3	4.76
	2	797.2	0.41	858.0	2.10	900.3	3.80	946.3	3.15
	3	797.4	0.52	869.1	3.02	924.9	4.63	976.4	2.25
	4	796.4*	0.45	855.4*	2.28	896.3	4.17	945.8	2.48
	5	797.2	0.45	858.0	2.64	906.0	4.27	965.2	2.31
	6	801.3	0.48	862.9	3.88	913.6	2.55	970.5	1.64
IX (2 & 3) *X and XI*	1	1110.4	1.17	1170.3*	4.99	646.8	5.77	540.0	5.0
	2			1170.3*	3.95	640.8*	5.47	519.3*	1.95
	3			1170.3*	11.44	645.8	1.24	539.5	2.32
	4			1182.3	4.19	669.6	2.47	602.5	2.08
	5			1182.3	2.90	650.9	3.42	596.7	2.26
	6			1204.3	9.42	650.9	2.87	596.7	1.43

* denotes optimal solution Z – solution reached t – CPU time in seconds on IBM 370/165

* There was an exception in seven solutions, where improvements were noted. These improvements were due to the present addition of the backtracking scheme, which was not utilized in our earlier work. Unfortunately, however, these improvements did not uncover any optimal solutions not previously found.

vided. The individual heuristic solution which matches the optimal solution as verified by other sources [3, 8, 13] is marked with an asterisk. The optimal solutions to problems which we added were verified with our experimental branch and bound code [10]. Due to the relatively high computation time requirements of the b & b code, we have not yet been able to obtain optimal solutions to problems VII–3, 4 and VIII–3, 4. The heuristic solutions for these problems are reported here to facilitate future comparisons. The optimal solutions to problems V–1 and VII–2, where the heuristic failed to find the optimal solution, are 1040.0 and 910.6, respectively. Those problems in which all the individual heuristic solutions are the same (indicated by an arrow, and in fact are optimal), are those in which all the warehouses were fixed, opened or closed due to the delta – omega cycle at the initial node.

VII. DISCUSSION OF THE RESULTS

The results look quite encouraging from the stand-point of both the "quality" of the solution and the computation times. The computer storage required on the other hand for solving these problems has been quite small (approximately 125K). It must be pointed out here that the low computing times are inspite of the fact that we have made no attempt to build into our code the efficiencies that can be derived by using the out-of-kilter transportation solution updating method. As the code exists now, whenever a solution is needed to a transportation problem, it is obtained as if it was an entirely new problem; in other words, without taking advantage of the solution to the transportation problem at the predecessor node.

Because of differences in computers and programming efficiences, a comparative evaluation of different solution procedures is difficult to make. We shall therefore simply identify the computational results reported on these same problems by other researchers. Our computation times for problem sets I–IV averaged 0.30 seconds, which includes times for the backtracking phase of the solution procedure. For these same problems, Sa fails to terminate with his b&b algorithm at the end of 15 minutes on IBM 360/75 and his heuristic solution times range from 1.55–9.00 minutes ([13], p. 1012). Ellwein reports solving only one such problem (I–1) with his exact code, which as reported took 3.27 minutes on IBM 360/67 ([3], p. 89). Of course, as the capacity constraints are not binding in these problems, they can much easily be solved using the uncapacitated codes.

Similar comparisons can be made on problem sets V, VI, VII–2, IX, X, and XI for computation times and solution quality. This comparison is, however, also limited by the fact that Sa and Ellwein do not report their results on all of these problems. The solution times with exact codes for these problems (excepting problem VII–2) are quite competitive. Problem VII–2 provided quite a challenge; Sa could not terminate his b&b algorithm within 15 minutes, however, his heuristic solution is better than ours which was reached in 2.66 minutes. For problems VII–3, 4 and VIII–3, 4, we were unable to obtain the optimal solutions within 2 minutes of CPU time with our experimental b&b code [10].

Analyzing the performance of the individual heuristics, we find that their computation times are not significantly different and their solutions are also reasonably "good". Of course, certain heuristics, particularly the largest omega and the smallest delta, do yield superior solutions. It should also be pointed out that the backtracking phase of the solution procedure improved the terminal solutions in several of these problems tested and is therefore highly recommended. The additional computation time due to the backtracking phase varied considerably among problems. Recall that this phase consists mainly of solving certain additional number of transportation problems and therefore its compu-

tation time varies directly with this number. Clearly, the computation times of this phase would be much improved with the incorporation of the out-of-kilter transportation solution updating method.

Lastly, to give an indication of the performance of the delta cycle including the two heuristic auxiliaries, we shall look at its application to problem X, which was most interesting in this respect. Problem X is one of Ellwein's problems and unlike the K & H problems, it is characterized by varying fixed costs and capacities and a sparse variable cost matrix, i.e., it contains a large number of prohibitive routes. For problem X, the application of the exact ΔV_i bounds at the initial node results in fix opening the first 5 out of the 15 free warehouses. When the $LB\Delta V_i$ bounds were applied, it opened only three warehouses (1, 2 and 5). The execution of the two heuristic auxiliaries then opened three additional warehouses in the following order of warehouse numbers: 4, 3, and 15. Warehouse number 15 was verified to be open in the optimal solution to problem X. For problem XI, the same two free warehouses were fixed opened by both the ΔV_i bounds and the $LB\Delta V_i$ bounds.

BIBLIOGRAPHY

[1] Davis, P. S. and T. L. Ray, "A Branch-Bound Algorithm for the Capacitated Facilities Location Problem," Nav. Log. Quart. Rev., Vol. 16 (Oct. 1969).

[2] Efroymson, M. A. and T. L. Ray, "A Branch-Bound Algorithm for Plant Location," Operations Research, Vol. 14 (May–June 1966).

[3] Ellwein, L. B., "Fixed Charge Location Allocation Problems with Capacity and Configuration Constraints," Technical Report No. 70–2, Department of Industrial engineering, Stanford University (Aug. 1970). Also, Ellwein, L. B. and P. Gray in AIIE Transactions, Vol. 3, No. 4 (Dec. 1971).

[4] Feldman, E., F. A. Lehrer, and T. L. Ray, "Warehouse Locations Under Continuous Economies of Scale," Management Science, Vol. 2 (May 1966).

[5] Geoffrion, A. M., The capacitated Plant Location Problem with Additional Constraints," Paper presented at the 42nd Joint National TIMS/ORSA/AIIE meetings, Nov. 1972.

[6] Gray, P., "Mixed Integer Programming Algorithm for Site Selection and Other Fixed Charge Problems," Department of Operations Research, Stanford University, Technical Report No. 6 (Nov. 1967).

[7] Jandy, G., "Approximate Algorithm for the Fixed Charge Capacitated Site Location Problem," Technical Report No. 67–3, Operations Research House, Stanford University (Apr. 1967).

[8] Khumawala, B. M., "An Efficient Branch and Bound Algorithm for the Warehouse Location Problem," Management Science, Vol. 18, No. 12 (Aug. 1972).

[9] Khumawala, B. M., "An Efficient Heuristic Procedure for the Uncapacitated Warehouse Location Problem," Nav. Res. Log. Quart. *20*, 109–121 (Mar. 1973).

[10] Khumawala, B. M. and U. Akinc; "An Efficient Branch and Bound Algorithm for the Capacitated Warehouse Location Problem," ORSA Bulletin on 43rd National Meeting (May 1973).

[11] Kuehn, A. A. and M. J. Hamburger, "A Heuristic Program for Locating Warehouses," Management Science, Vol. 9 (July 1963).

[12] Revelle, C., D. Marks, and J. C. Liebman, "An Analysis of Private and Public Sector Location Models," Management Science, Vol. 16 (July 1970).

[13] Sa, G., "Branch and Bound and Approximate Solutions to the Capacitated Plant Location Problem," Operations Research (Nov.–Dec. 1969).

14] Spielberg, K., "An Algorithm for the Simple Plant Location Problem with Some Side Conditions," Operations Research *17*, 81–111 (1969).

15] Spielberg, K., "Plant Location with Generalized Search Origin," Management Science, Vol. 15 (Nov. 1969).

AN ALGORITHM FOR THE SOLUTION OF A LOCATION PROBLEM WITH METRIC CONSTRAINTS

Margaret K. Schaefer

University of Illinois, Chicago Circle
Chicago, Illinois

and

Arthur P. Hurter, Jr.

Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, Illinois

ABSTRACT

In many location problems, the solution is constrained to lie within a closed set. In this paper, optimal solutions to a special type of constrained location problem are characterized. In particular, the location problem with the solution constrained to be within a maximum distance of each demand point is considered, and an algorithm for its solution is developed and discussed.

1. INTRODUCTION

The problem concerned with finding the coordinates of the point in R^m which minimizes the sum of the weighted distances from that point to a set of known points is often referred to as the Weber problem.

The problem, denoted (PB), can be written:

$$\text{(PB):} \qquad \min_x f(x) = \sum_{i=1}^{n} w_i d(a_i, x)\,,$$

where $\{a_i, i=1, \ldots, n\}$ is a finite set of demand points with known coordinates, the $\{w_i\}$ are non-negative scalar weights, and $d(a_i, x)$ is any metric.

Extensive work has been done on this problem, both for the case where distance is measured as Euclidean distance [1], [2], where it is measured as rectilinear distance [15], and for more general metrics, [16]. Several efficient algorithms exist to solve the Weber problem using both kinds of distance measures [1], [7], [10].

The Weber problem with metric constraints can be written:

$$\text{(PA):} \qquad \min_x f(x) = \sum_{i=1}^{n} w_i d(a_i, x); \qquad \text{s.t. } d(a_i, x) \leq d_i^{\#}, \qquad i=1, \ldots n.$$

A special case is of particular interest:

$$(\mathrm{PA}^{\#}): \qquad \min_{x} f(x) = \sum_{i=1}^{n} w_i d(a_i, x); \qquad \text{s.t. } d(a_i, x) \leq d^{\#}.$$

(PA$^{\#}$) is the special case of (PA) in which the point x is constrained to be within the same distance, $d^{\#}$, of the demand points, a_i (i.e., $d_i^{\#} = d_j^{\#} = d^{\#}$ all $i, j = 1, \ldots n$).

Solutions to the Weber problem with metric constraints are described, conditions under which a solution will exist are explored, and a solution algorithm is developed in this paper. Toregas and Revelle [13] have worked on a similar problem.

Problems (PA) and (PA$^{\#}$) are of interest as location problems in cases where minimization of the sum of the weighted distances does not adequately reflect the true objective. For example, in locating a fire station, it is generally accepted that response time is very important in determining damage [3], [12]. To minimize damage, the problem might be formulated as a Weber problem where the sum of the weighted distances to the fire station is to be minimized through selection of an optimal fire station location. However, since fire departments must protect lives as well as property, an upper bound on the allowable distance from each neighborhood to the station is appropriate as a life safety-measure. This problem can be represented in the forms (PA) or (PA$^{\#}$), where the demand points may be neighborhoods and the weights the expected annual dollar damage by fire in the neighborhood, i.e., the total value at stake multiplied by the probability of a fire.

Other examples may be cited. All other emergency services such as police and ambulance services face analogous problems. This formulation could also be used to locate schools, hospitals, retail stores, banks, and warehouses, where an upper bound on distance to customers is important. Although most of these examples refer to location on a plane, the theorems below apply to m dimensional space.

2. NOTATION

Let $\{a_1, a_2, \ldots, a_n\}$ be any set of n demand points in R^m.

Nonnegative scalar weights, $\{w_i\}$, are associated with the demand points.

Let $H \underline{\underline{\Delta}}$ the convex hull of $\{a_i\}$.

Let $\{a_j^{\#}\}$ be the set of extreme points of H; let I denote the index set of the extreme points. Thus, $\{a_j^{\#}\} = \{a_i | i \epsilon I\}$.

Let $S \underline{\underline{\Delta}} \{x | d(x, a_i) \leq d_i^{\#}, i = 1, \ldots, n\}$, the feasible set for (PA) and (PA$^{\#}$).

Let $S' \underline{\underline{\Delta}} \{x | d(x, a_j^{\#}) \leq d_j^{\#}, j \epsilon I\}$.

Let x_B^*, x_A^* and $x_{A\#}^*$ denote the optimal solutions to (PB), (PA), and (PA$^{\#}$), respectively.

Let $U_i \underline{\underline{\Delta}} \{x | d(x, a_i) = d_i^{\#}\}$. U_i is the hypersphere around a_i of radius $d_i^{\#}$.

3. DEFINITIONS AND A THEOREM FOR REFERENCE

A *constraining point* of S is a demand point a_i such that U_i forms part of the boundary of S.

An *intersection point* is a point x^0 such that both $d(x^0, a_i) = d_i^{\#}$ and $d(x^0, a_j) = d_j^{\#}$ for two distinct demand points a_i and a_j. x^0 occurs at the intersection of the ith constraining arc U_i and the jth, U_j.

A point $x_1 \epsilon S$ is said to be *visible* from a point $x_2 \epsilon S$ if the straight line segment connecting the two points contains no points of S except x_1. A point $x_1 \epsilon S$ is said to be visible from a set C, where $C \cap S = \varphi$

if x_1 is visible from some point on C. This concept was first used by Goldman (see Witzgall [16]). The concept of visibility has resulted [4] in the theorem (Visibility Theorem):

If x_B^* is not feasible for (PA), there exists an optimal solution to (PA) which is visible from x_B^*.

For (PA), in two dimensions, the theorem is illustrated in Figure 1. x_B^* is the optimal solution to unconstrained (PB). If x_B^* is not in S, then the optimal solution to (PA) is visible from x_B^* as illustrated. Thus, an optimal solution to (PA) can be found among the points on the boundary of S which are visible from x_B^*.

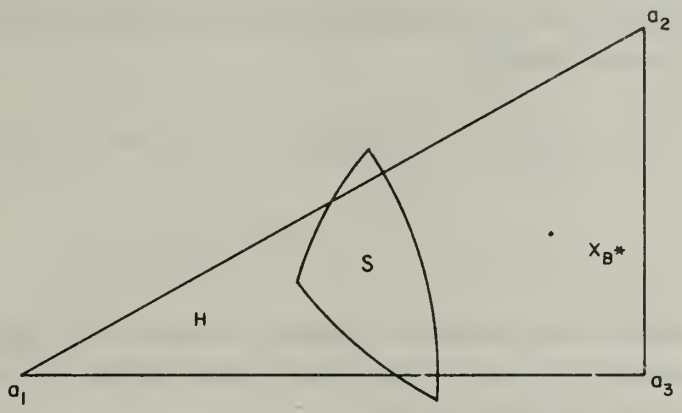


FIGURE 1.

The influence of changes in the weights, w_i, on the value of the objective function of (PB) is summarized in Theorem 1.

THEOREM 1: The objective function of (PB), $f(x)$, is strictly increasing in each weight w_i except when $x_B^* = a_i$, in which case $f(x_B^*)$ remains constant as w_i changes.

PROOF: Assume $x_B^* \neq a_i$. Let x_B^* be the optimum for (PB) with weights w_i. Let $w_i' = w_i - \epsilon$ for any $\epsilon > 0$. Then $w_i' d(a_i, x_B^*) < w_i d(a_i, x_B^*)$. Since $w_j' = w_j$, $j = 1, \ldots, n$; $j \neq i$, $\sum_{j=1}^{n} w_j' d(a_j, x_B^*) < \sum_{j=1}^{n} w_j d(a_j, x_B^*) = f(x_B^*)$. Reoptimization of (PB) using w_i' can only further reduce the value of the objective function.

4. RESULTS CHARACTERIZING S

THEOREM 2: For (PA), a necessary condition for $S \neq \varphi$ is that $d(a_j^\#, a_k^\#) \leq d_j^\# + d_k^\#$ for $j, k \in I$.

PROOF: To prove the theorem, we assume the opposite and show a contradiction. Assume $\exists a_j^\#$ and $a_k^\# \ni d(a_j^\#, a_k^\#) > d_j^\# + d_k^\#$ and that $S \neq \varphi$. Then $d(a_j^\#, a_k^\#) = d_j^\# + d_k^\# + \epsilon$ for some $\epsilon > 0$. Let x^0 be any point in S. Then from the definition of S, $d(a_i^\#, x^0) \leq d_i^\#$, $i = 1, \ldots, n$. Now, from the Triangle Inequality, $d(a_j^\#, a_k^\#) \leq d(a_j^\#, x^0) + d(a_k^\#, x^0)$ so $d_j^\# + d_k^\# + \epsilon \leq d(a_j^\#, x^0) + d(a_k^\#, x^0)$. Since $d(a_j^\#, x^0) \leq d_j^\#$ and $d(a_k^\#, x^0) \leq d_k^\#$, we have $d_j^\# + d_k^\# + \epsilon \leq d_j^\# + d_k^\#$. This is impossible since ϵ is positive. Thus, there exists no such point x^0 in S, and the theorem has been proved.

THEOREM 3: For (PA#), assume $S \neq \varphi$. If for any extreme point $a_j^\#$, $\max d(a_i^\#, a_j^\#) \leq d^\#$, then $a_j^\#$ is not a constraining point of S.

PROOF: Since max $d(a_i^{\#}, a_j^{\#}) \leq d^{\#}$, the set $S = \{x | d(a_j^{\#}, x) \leq d^{\#}, i = 1, \ldots, n\}$ includes $a_j^{\#}$. Therefore, the set $S_j = \{x | d(a_j^{\#}, x) \leq d^{\#}\}$ strictly contains S. Thus, it cannot be a constraining point of S.

This theorem is true for (PA), if the objective is changed to $\max_i d(a_i^{\#}, a_j^{\#}) \leq \min_i d_i^{\#}$. For (PA#), if $d(a_j^{\#}, a_i^{\#}) \leq d^{\#}$ for an extreme point a_j and all $i \epsilon I$, then a_j can be excluded from I. Thus, I is now reduced to a subset I' of possibly constraining extreme points.

5. LAGRANGEAN INTERPRETATIONS

From now on, assume that the same metric appears in the objective and in the constraint so (PA) can be written in Lagrangean form as:

(LPA):

$$\min_{(x,\lambda)} L(x, \lambda) = \Sigma w_i d(a_i, x) + \Sigma \lambda_i [d(a_i, x) - d_i^{\#}]$$

$$= \Sigma (w_i + \lambda_i)\, d\,(a_i, x) - \Sigma \lambda_i d_i^{\#},$$

where $\lambda_i \Delta$ the optimal Kuhn-Tucker multiplier for the ith constraint of (PA). Since Slater's constraint qualification holds in all nontrivial cases, the existence of such a multiplier is guaranteed by the Kuhn-Tucker saddlepoint necessary optimality theorem.

The similarity between (PB) and (LPA) suggests that techniques for solving (PB) can be adapted for solution of the constrained problem (PA). The multipliers also yield economic insight into the problem. For (PA), λ_i is the amount that w_i would have to be increased to make x_A^* an unconstrained optimal solution to the Weber problem, (PB). In effect, λ_i is an artificial weight added to the original weight w_i in order to "pull" the unconstrained optimum into the feasible region, S, for (PA). The solution technique presented later is based upon this interpretation. The development in this section depends upon the similarity in the way the constraint and the objective appear in the Lagrangean function.

As $d^{\#}$ increases, the problem becomes less constrained and S increases in size. The new objective function value will decrease toward the unconstrained minimum. As $d^{\#}$ increases, the additional "weights" λ_i will decrease. Thus, the optimal objective function value decreases strictly as the weights decrease. Therefore, Theorem 1 holds for (PA) as well as (PB).

6. OVERVIEW OF ALGORITHM TO SOLVE (PA) IN TWO DIMENSIONS

The algorithm is based upon the Lagrangean form of (PA), (LPA). It involves solving a series of unconstrained Weber problems (PB). An algorithm for the solution to the Weber problem is available, but it is not exact [1]. If this algorithm is used, the effect of its errors on the present algorithm are unknown. An exact, efficient solution to the Weber problem in m-space has been developed [10] and may be used in conjunction with the present algorithm. An algorithm for the solution of constrained Weber problems in m-space is also available [5], but is more cumbersone than the present algorithm for problem (PA) in two dimensions.

In any infinite algorithm it is not possible to get the exact optimal solution, but only to approach it to the required degree. The required approach is defined in terms of the constraints $d(x, a_i) \leq d_i^{\#}$. The algorithm will terminate when an optimal value for the objective has been found such that the optimal location is within $d_i^{\#} + \epsilon_j$ of each demand point a_i.

Assume $S \neq \phi$. (Theorem 2 provides a simple way to eliminate some $S = \phi$ cases.) The unconstrained problem (PB) is solved using one of the several existing algorithms and the optimal point is denoted by x_B^*. If $x_B^* \epsilon S$, x_B^* solves (PA). If x_B^* is not feasible for (PA), let F be the portion of S that is visible from x_B^*. By the Visibility Theorem we know that x_A^* will be located on F. Let J be the set of points whose constraints are violated by x_B^*; that is, $J \triangleq \{a_j | d(x_B^*, a_j) > d_j^{\#}\}$. In two dimensions a typical problem is illustrated in Figure 2.

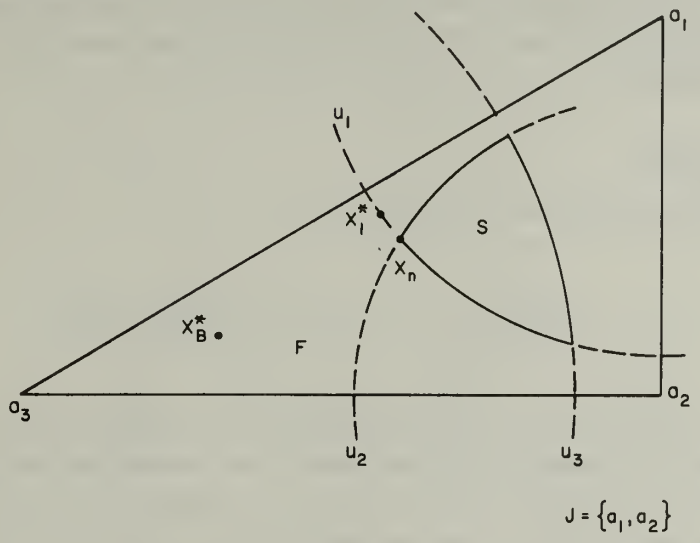


FIGURE 2.

Consider the following variation of (PA):

$$(\mathrm{PA}_i): \qquad \min_x \sum_{j=1}^{n} w_j d(a_j, x) \qquad \text{s.t. } d(a_i, x) = d_i^{\#} \qquad \text{for point } a_i \text{ only.}$$

Equivalently, in Lagrangean form:

$$(\mathrm{PA}_i): \qquad \min_{x,\lambda} \sum_{j=1}^{n} w_j d(a_j, x) + \lambda_i (d(a_i, x) - d_i^{\#}).$$

The algorithm involves solving (PA_i) for each point $a_i \epsilon J$. The constraint insures that the solution x_i^* to (PA_i) is located on U_i. The Visibility Theorem requires that x_i^*, if it is a feasible solution for (PA), be located on the border of S visible from x_B^*.

Solving (PA_k) for each $a_k \epsilon J$ involves finding the point x_k^*. x_k^* is the solution to (PB) with the weight on a_k increased so that x_k^* is on U_k, all other weights remaining the same. The weight added to w_k in order to "pull" the unconstrained optimum to a point on U_k, x_k^*, is denoted λ_k^*.

If x_k^* is feasible, it could be the optimum for (PA), x_A^*. In the algorithm, $f(x_k)$ is compared with an upper bound f (x'), found by evaluating $f(x)$ at the intersection points of S, or with the $f(x_j^*)$ obtained from previously solved problem (PA_j). If $f(x_k^*)$ is less than the previous upper bound, it becomes the new upper bound in the algorithm and is then denoted UB.

This procedure is continued until all $a_k \epsilon J$ are examined. The upper bound UB at the end of the algorithm is the optimum, $f(x_A^*)$. It may not be necessary to solve (PA_k) for some a_k's because of a simple check (step 3 of the algorithm). If in the nth step of the solution of (PA_k), the objective function $f(x_n)$ is greater than the current upper bound *and* the next step would be to increase the weight λ_k, then (PA_k) can be abandoned, because x_k^* will not be the optimum x_A^* (see Theorem 6 below).

Prior to presentation of the algorithm the two major features, solution of the problems (PA_i) and the evaluation and identification of the intersection points, are briefly discussed. In addition, three theorems concerned with properties of the objective function are presented since they are used in developing the algorithm.

a. The Solution of (PA_i)

Notice that (PA_i) can be written in the form

$$(PA_i): \qquad \min_{x,\lambda} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_j d(a_j, x) + (w_i + \lambda_i)[d(a_i, x) - d_i^{\#}].$$

An increase in λ_i is equivalent to an increase in the weight w_i associated with the point a_i. Theorem 1 insures that the objective function strictly increases as any weight increases. Thus, to solve (PA_i) it is necessary to find the smallest $\lambda_i > 0$ such that $x_{\lambda_i}^*$, the solution to (PB) using the weight $w_i + \lambda_i$, all other weights unchanged, is feasible with respect to a_i. This corresponds to a search over λ_i and is the basis of the algorithm. Since the search procedure operates on one demand point and its constraint at a time, the intersection points are separately considered.

b. Separate Determination of the Intersection Points

It is necessary to determine the feasible intersection points and to evaluate the objective function at them. The feasible intersection point with the lowest value of the objective function serves as an upper bound on the minimization in (PA). These intersection points labeled $x_\cap$ on Figure 2, can be readily determined. In R^3 there are no unique points at which to evaluate the objective as there are in R^2, since there are then intersection arcs. However, the intersection points are used only to find an upper bound $f(x')$ in the algorithm and, therefore, are not strictly necessary. In R^3 or spaces of higher dimensions, the value of the objective at the first feasible point generated by the algorithm could serve as an initial upper bound which is constantly revised as the solution procedure progresses.

For many a_i, a_j choices, intersection points will not exist as illustrated on Figures 3(a) and (b). Figure 3(a) corresponds to $d(a_i, a_j) > d_i^{\#} + d_j^{\#}$. Whenever a pair of demand points (a_i, a_j) exists such that $d(a_i, a_j) > d_i^{\#} + d_j^{\#}$, then $S = \phi$ by Theorem 2. If $d(a_i, a_j) = d_i^{\#} + d_j^{\#}$, for any pair a_i, a_j, the feasible region S is reduced in dimension. For example, in R^2, S is reduced to a point if the Euclidean metric is used and to a point on a line if the rectilinear metric is used. The algorithm will not work in this case and some other method must be used. In two dimensions the required alteration appears straightforward. The case depicted in Figure 3(b) occurs when the circle of radius $d_i^{\#}$ around one demand point (a_i on the figure) is contained in the circle of radius $d_j^{\#}$ around another demand point (a_j on the figure). In this case there is no intersection point, but S is not necessarily empty. These simple checks can be applied

immediately and will indicate whether $S=\varphi$ and perhaps eliminate many of the $\frac{r(r-1)}{2}$ (a_i, a_j) combinations, where r is the number of points in J.

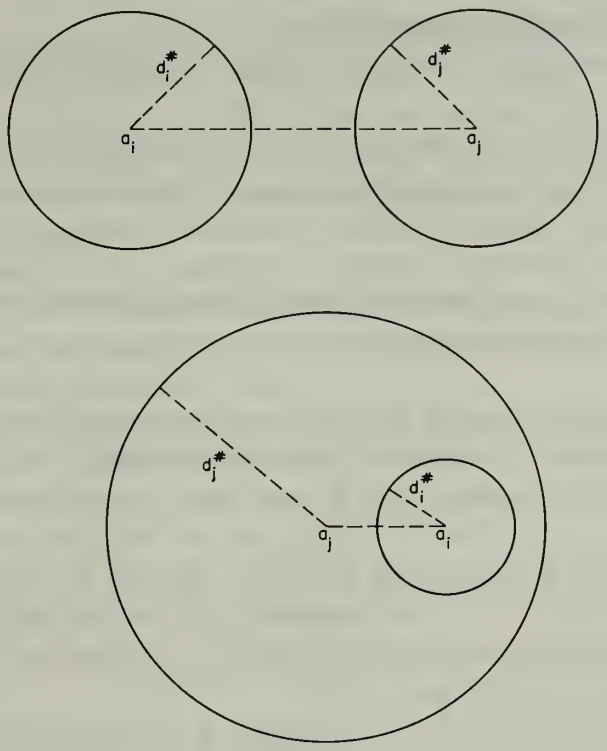


FIGURE 3.

The existing intersection points can be found by calling a simple computer subroutine for solving systems of equations. For the rectilinear metric the intersection of two lines is determined. In two dimensions, the feasible set, S, could be a line with an infinite number of intersections generated by two demand point constraints. Here, the ends of the line are the intersection points to be evaluated. Intersections of more than two constraints will appear more than once when demand points are considered pairwise, and hence, will cause no problem.

The intersection points must be feasible to be potential solutions for (PA). Thus, the next step is to test the intersection points for feasibility. Infeasible intersection points are eliminated from further consideration.

At the end of this preliminary procedure, the objective function value associated with each feasible intersection point is evaluated. Let x' denote the intersection point with lowest $f(x)$ value. Then $f(x')$ is an upper bound for (PA) since x' is feasible. In the algorithm, $f(x')$ is the initial value of UB. As the algorithm progresses, improved values for UB are found.

c. Three Theorems Necessary for the Algorithm

It is well known that a sufficient condition for x_B^* to be located at a_k is $w_k \geq \sum_{\substack{i=1 \\ i \neq k}}^{n} w_i$. Thus, we have an

upper bound on λ_k for any constraining point a_k, namely, $\lambda_k^{\#} = \sum_{\substack{i=1 \\ i \neq k}}^{n} w_i - w_k$. A Bolzano search on λ over the interval $[0, \lambda_k^{\#}]$ will yield the optimal λ_k^*. The associated optimal location $x_{\lambda_k}^*$ is on U_k, so $x_{\lambda_k}^* = x_k^*$. Let $g_i(\lambda_i) = d(x_{\lambda_i}, a_i)$, $a_i \epsilon J$, where x_{λ_i} is the optimal point found by solving (PB) with $w_i' = w_i + \lambda_i$, the rest of the weights remaining the same.

THEOREM 4: Assume $x_B^* \neq a_i$. Let $g_i(\lambda_i) = d(x_{\lambda_i}, a_i)$; then for $d(x_{\lambda_i}, a_i) = d' > 0$, $g_i(\lambda_i)$ is strictly decreasing.

PROOF: When $\lambda_i = 0$, $x_{\lambda_i} = x_B^*$ and $d(x_B^*, a_i) > 0$ because $x_B^* \neq a_i$. Therefore $g_i(0) > 0$. Also we know that $g_i(\lambda_i^{\#}) = 0$, so $g_i(\lambda_i)$ is eventually decreasing. To show it is strictly decreasing, for positive $g_i(\lambda_i)$, assume the opposite. Assume that there are two positive scalars λ^0 and λ^1, $\lambda^0 > \lambda^1$, such that for $d' > 0$, $g_i(\lambda^0) = d'$ and $g_i(\lambda^1) = d'$, so $d(x_{\lambda^0}, a_i) = d(x_{\lambda^1}, a_i) = d'$. Then λ^0 and λ^1 both are Lagrange multipliers for (PA_i) using d' as $d_i^{\#}$ and $f(x_{\lambda^0}) = f(x_{\lambda^1})$. But $\lambda^0 > \lambda^1$ so $f(x_{\lambda^0}) > f(x_{\lambda^1})$. There is a contradiction. Consequently there cannot exist two distinct λ_i's for any positive d' and $g_i(\lambda_i)$ is strictly decreasing.

Let K_j be the feasible set of points on U_j which are *not* intersection points so $K_j = \{x \epsilon S | d(x, a_j) = d_j^{\#}, d(x, a_i) \neq d_i^{\#}, i \neq j, i = 1, \ldots, n\}$. Then x_j^* which solves $(PA_j) \epsilon K_j$.

THEOREM 5: If x_j^* is not feasible, $x_A^* \notin K_j$. In other words, x_A^* is not on the portion of S formed by the a_j constraint alone.

PROOF: For a point in K_j, only λ_j can be positive. Because of complementary slackness, at x_j^*, $\lambda_j^* > 0$, $\lambda_i^* = 0$, $i \neq j$, $i = 1, \ldots, n$. By Theorem 4, λ_j^* is the only multiplier associated with $d_j^{\#}$. Thus it is not possible to reach another point on K_j by increasing or decreasing λ_j. Therefore, if x_j^* is not feasible, the optimal point x_A^* cannot be in K_j.

Let ${}_n\lambda_k$ denote the nth iteration value of λ_k, as in step 1 of the following algorithm. Let x^0 denote the feasible point which leads to the current upper bound, $UB = f(x^0)$.

THEOREM 6: If both $g_k({}_n\lambda_k) > d_k^{\#}$ and $f(x_{n\lambda_k}) > UB = f(x^0)$, the optimal point x_A^* is not characterized by $\lambda_k > 0$, all other $\lambda_i = 0$. In this case, (PA_k) need not be solved since it cannot yield the optimum.

PROOF: Theorem 1 indicates that $f(x_\lambda)$ increases monotonically as λ increases, because increasing λ is equivalent to increasing a weight. Since $g({}_n\lambda_k) > d_k^{\#}, \lambda_k$ would be increased in the Bolzano search of Step 4 in the following algorithm. Thus, $f(x_{n+1\lambda_k}) > f(x_{n\lambda_k}) > f(x^0)$. Since $f(x)$ is convex, the feasible point x^0 will always be better than any $x_{m\lambda_k}$, where ${}_m\lambda_k > {}_n\lambda_k$. Thus, a better feasible point than x^0 cannot be found by changing λ_k for point a_k.

7. ALGORITHM TO SOLVE (PA)

0. Solve (PB) for x_B^*. This defines set J, the demand points with constraints violated by x_B^*. If $J = \phi$, x_B^* solves (PA). Order the points in J, $\{a_1, a_2, \ldots, a_k, \ldots, a_r\}$ arbitrarily. Determine the intersection points in a pairwise fashion. Check them for feasibility with respect to all constraints and evaluate $f(x_n)$ for those that are feasible. Determine x', the intersection point which minimizes the objective function, and set $UB = f(x')$. If no feasible intersection points exist, $S = \varphi$. Consider a_1 first, so $k = 1$.

1. Let n denote the iteration and set $n = 1$ initially. Let $a_k \epsilon J$ be the demand point under consideration and ${}_1\lambda_k = \lambda_k^{\#}/2$. Let ${}_i\lambda_k$ denote the additional weight λ added to w_k in the ith iteration.

2. Find the unconstrained optimum $x_{n\lambda_k}$ using $w_k' = w_k + {}_n\lambda_k$, all other weights being the given w_i, in the (PB) algorithm. Calculate $g_k({}_n\lambda_k) - d_k^{\#}$. Call it Q_k.

3. If both $Q_k > 0$ and $f(x_{n\lambda_k}) > UB$, stop and go to a new $a_k \epsilon J$ in Step 1 (by Theorem 6). If not go to step 4.

4. Then if $Q_k > \epsilon_k$, go to step 2 with new λ_k, ${}_{n+1}\lambda_k = {}_n\lambda_k + \lambda_k^{\#}/2^{n+1}$; if $Q_k < -\epsilon_k$, go to step 2 with new λ_k, ${}_{n+1}\lambda_k = {}_n\lambda_k - \lambda_k^{\#}/2^{n+1}$; if $|Q_k| \leq \epsilon_k$, let $\lambda_k^* = {}_n\lambda_k$ and go to step 5.

5. If x_k^* is feasible for (PA), $f(x_{\lambda_k}^*) < UB$ since we were not shut off by step 3. So $f(x_{\lambda_k}^*)$ becomes the new upper bound, UB. Then go to step 1 with $a_{k+1} \epsilon J$. If $x_{\lambda_k}^*$ is not feasible for (PA), by Theorem 5 we can immediately go to step 1 with a_{k+1}. The optimal solution is the upper bound UB left when all the points in J have been considered.

8. SIMPLIFICATIONS IN THE ALGORITHM FOR (PA$^{\#}$)

Several simplifications are possible for (PA) with the help of Theorem 7.

THEOREM 7: For (PA$^{\#}$), $S \cap H = S' \cap H$.

PROOF: Let x^0 be any point in $S' \cap H$. Because H is convex and bounded, $d(x, x^0)$ achieves its maximum at an extreme point of H. Thus, $d(a_j^{\#}, x^0) \leq d^{\#}$, $\forall_j \epsilon I$ because $x^0 \epsilon S'$. $d(x, x^0) \leq d^{\#}$, $\forall x \epsilon S \cap H$. Since $\{a_i\} \subset H$, $d(a_i, x^0) \leq d^{\#}$, $i = 1, \ldots, n$, so $x^0 \epsilon S \cap H$. Thus $(S' \cap H) \subseteq (S \cap H)$.

Conversely, suppose $x^0 \epsilon S \cap H$, so that $d(a_i, x^0) \leq d^{\#}$, $i = 1, \ldots, n$. Then, since $\{a_j\}$, $\subseteq \{A_i\}$, $d(x^0, a_j^{\#}) \leq d^{\#}$, $j \epsilon I$. So $(S \cap H) \subseteq (S' \cap H)$. Therefore, $S' \cap H = S \cap H$.

Theorem 7 is not true for (PA). Consider a case where $S' \neq \varphi$, but $\exists$ demand points a_i and a_j, with the property that $d_i^{\#} + d_j^{\#} < d(a_i, a_j)$. Then $S' \neq \varphi$ but by Theorem 1, $S = \varphi$.

In the cases when $d(\cdot, \cdot)$ is Euclidean or when the problem is in two dimensions with any norm [14], $x_{\lambda}^{\#} \epsilon H$. In those cases, Theorem 7 is useful for (PA$^{\#}$). In order to generate the feasible solution set S for (PA$^{\#}$), only those points which are within $d^{\#}$ of the extreme points, $a_j^{\#}$, must be considered. This property will, in general, make S easier to determine. In particular, in (PA$^{\#}$), the constraints can now be written subject to $d(x, a_j^{\#}) \leq d^{\#}$, $j \epsilon I$.

When solving (PA$^{\#}$), not all the possible intersection points must be found. Only the extreme points of the convex hull can be constraining points of S for (PA$^{\#}$), by Theorem 7. Thus, the number of intersection points is considerably reduced for most problems to the feasible intersection points formed by the intersection of the constraints of pairs of extreme points of the convex hull. In addition, some of the extreme points may be eliminated as possible constraining points by Theorem 3.

9. CONVERGENCE OF THE ALGORITHM

For each point $a_i \epsilon J$ we want to find the point x_i^* which solves (PA$_i$):

$$(\text{PA}_i): \qquad \min_x f(x) = \sum_{i=1}^{n} w_i d(a_i, x) \qquad \text{s.t. } d(a_i, x) = d_i^{\#}$$

$l_i) = d(x_{\ l}$

From Theorem 4 we know that $g_i(\lambda_i) = d(x_{\lambda i}, a_i)$ is monotone decreasing. Thus, $\exists$ a unique λ_i^* such that $g_i(\lambda_i^*) = d_i^{\#}$ and if $g_i(\lambda_i') \geq d_i^{\#}$, $\lambda_i' \leq \lambda_i^*$. Thus, the one-dimensional Bolzano search for λ_i^* over the interval $[0, \lambda_i^{\#}]$ clearly converges to λ_i^*, since $g_i(\lambda_i)$ converges on $d_i^{\#}$ by following the rule in step 4 of the algorithm. Thus, the solution to (PA) is always found.

An upper bound on the number of iterations necessary to get within δ_i of the optimal λ_i^* can be found. At the nth iteration the Bolzano search reduces the interval containing λ_i^* to one of size $\lambda_i^{\#}/2^n$. Therefore, to get within δ_i of λ_i^*, it is sufficient to go through n' iterations, where n' is the smallest positive integer satisfying $\delta_i \geq \lambda^{\#}/2^n$, or equivalently, $n \geq \dfrac{\log \lambda_{\#} - \log \delta_i}{\log 2}$. Using this relationship, the following values for n_i are required for the associated ratios of δ_i and $\lambda_i^{\#}$.

Typical values

$\delta_{i/\lambda_i^{\#}}$	n_i	$\lambda_i^{\#}$	δ_i
1/100	7	1	0.01
1/200	8	2	0.01
1/400	9	4	0.01
1/1000	10	1	0.001

Therefore, increased "accuracy" seems obtainable at little additional effort measured by iterations required. The values of δ_i can be employed in a stopping rule for termination of this phase of the algorithm. For example, terminate when $|{}_i\lambda_k - {}_{i+1}\lambda_k| < \delta_k$.

The algorithm proceeds by determining the demand points that may be constraining points. Choose one such demand point, a_i. Now the problem of importance is (PA$_i$), a Weber problem with a single norm constraint. The algorithm has been shown to converge on the correct solution for each (PA$_i$), so it is obvious that the overall algorithm must find the correct solution.

However, since the algorithm is made up of a series of (PB) problems, for which only approximate algorithms may be used, x_A^* will be an approximate solution.

10. COMPUTATIONAL RESULTS OF THE ALGORITHM

The algorithm has been programmed in Fortran IV for use on the CDC 6400 at Northwestern University. The results of test runs of several sample problems are given in the appendix.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of the National Science Foundation through Grant Number GU–3851 to the Urban Systems Engineering Center, The Technological Institute, Northwestern University and through Grant Number GK–38336.

The authors also wish to express their appreciation to Professor Roy Marsten, Department of Industrial Engineering and Management Sciences, Northwestern University for his help in developing an efficient program for problem (PA).

APPENDIX

Computational Results Using the Algorithm for Solution of Location Problems with Metric Constraints

The algorithm to solve the Weber problem with metric constraints was solved for several test problems using Euclidean distance on a plane. The following results were obtained on the CDC 6400 computer at Northwestern University. All the problems were run separately; savings in computer time could be realized by running a series of these problems simultaneously.

Problems	a_i	w_i	$d_i^{\#}$	Optimum	Required computer time (sec)	Remarks
1	(0, 0) (0, 1) (2, 0.5)	1.0 1.0 3.0	1.0 1.0 2.0	$x_B^* = (2, 0.5)$ $J = \{1.2\}$ $x_A^* = (0.866, 0.500)$ Intersection of U_1 and U_2 so $\lambda_1^* > 0$, $\lambda_2^* > 0$. $f(x_A^*) = 6.40$.	8.6	By varying λ_1 and then λ_2 separately, points were achieved on U_1 and U_2, respectively but they were infeasible. By Theorem 6, $x' = (0.866, 0.500)$ is optimal.
2	(0, 0) (0, 1) (2, 0.5)	2.0 1.0 3.0	1.0 1.0 2.0	$x_B^* = (2, 0.5)$ $J = \{1, 2\}$ $x_A^* = (0.834, 0.460)$ $\lambda_1^* = 0.488$ $f(x_A^*) = 6.39$	8.9	Increasing the weight on a_1, from $w_1 = 1$ in problem 1 to $w_1 = 2$ here, changes the optimum.
3	(0, 0) (3, 2) (5, 5)	1.0 1.0 1.0	1.0 2.0 10.0	Optimum doesn't exist, since $S = \phi$.	2.35	(0, 0) and (3, 2) have no intersection points, as in Fig. 3(a). Program printed, "The feasible set is empty."
4	(0.0) (0, 1) (1, 0) (5, 8)	2.0 1.0 1.0 1.0	2.0 2.0 2.0 8.0	$x_B^* = (0.177, 0.227)$ $J = \{4\}$ $x_A^* = (0.764, 1.22)$ $f(x_A^*) = 12.904$	10.8	
5	(−2, 2) (−3, −1) (0, 0) (1, 1) (−1, −2) (3, 0) (1.5, 0.5) (4, −1)	1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0	5.0 6.0 3.0 6.0 6.0 6.0 2.0 3.0	$x_B^* = (0.478, -0.065)$ $J = \{8\}$ $x_A^* = (1.11, -0.223)$ $\lambda_3^* = 0.882$ $f(x_A^*) = 18.49$	17.9	

REFERENCES

[1] Cooper, L., "Location-Allocation Problems", Operations Research, Vol. 2 (1963).

[2] Francis, R. L. and A. V. Cabot, "Properties of a Multifacility Location Problem Involving Euclidean Distance", Nav. Res. Log. Quart., *19*, 335–355 (June 1972).

[3] Hogg, J., "The Siting of Fire Stations", Operational Research Quarterly, *19* (1968).

[4] Hurter, A. P., M. K. Schaefer, and R. E. Wendell, "Solutions of Constrained Location Problems", to appear, *Management Science*.

[5] Hurter, A. P., and Alejo Planchart, "Constrained Multifacility Location Problems with Mixed Norms: A Geometric Programming Approach", to appear, SIAM J. of Control.

[6] Kuhn, H. "Location Problems and Mathematical Programming", in Special 1963 Colloquium on the Application of Mathematics to Economics (Budapest, Publishing House of the Hungarian Academy of Sciences, 1965) pp. 235–242.

[7] Kuhn, H. and R. Kuenne, "An Efficient Algorithm for the Numerical Solutions of the Generalized Weber Problem in Spatial Economics" J. of Regional Science, *4* (1962).

[8] Love, R. F., "Locating Facilities in Three-Dimensional Space by Convex Programming", Nav. Res. Log. Quart., *16*, 503–516 (Dec. 1969).

[9] Mangasarian, O. L., *Nonlinear Programming*, (McGraw-Hill Publishing Company, New York, 1969).

[10] Planchart, A. and A. P. Hurter, "An Efficient Algorithm for the Solution of the Weber Problem with Mixed Norms", to appear in *SIAM Journal of Control*.

[11] Planchart, Alejo, "A Geometric Programming Approach to the Problem of Finding Efficient Algorithms for Constrained Location Problems with Mixed Norms", N.U. Ph. D. Dissertation, June 1973.

[12] Toregas, C., R. Swain, C. Revelle and L. Bergman, "The Location of Emergency Services", Operations Research, *19* (Nov. 1971).

[13] Toregas, C. and C. Revelle, "Optimal Location under Time or Distance Constraints", Papers of the Regional Science Association, Vol. 28.

[14] Wendell, R. E. and A. P. Hurter, "Location Theory-Dominance and Convexity," Operations Research, *21*, 314–321 (Jan.-Feb. 1973).

[15] Wesolowsky, G. O. and R. F. Love, "Location of Facilities with Rectangular Distances among Point and Area Destinations", Nav. Res. Log. Quart., March 1971.

[16] Witzgall, C., "Optimal Location of a Central Facility: Mathematical Models and Concepts", National Bureau of Standards Report 8388, June 1965.

EFFICIENT COMPUTATIONAL DEVICES FOR THE CAPACITATED TRANSPORTATION PROBLEM

Robert W. Langley
United States Air Force Academy

Jeff Kennington
Southern Methodist University

C. M. Shetty
Georgia Institute of Technology

ABSTRACT

This paper presents the details for applying and specializing the work of Ellis Johnson [10] and [11] to develop a primal code for the well-known capacitated transportation problem. The code was developed directly from the work of Johnson, but is similar to codes developed by Glover, Karney, Klingman, and Napier [6] and Srinivasan and Thompson [14]. The emphasis in the presentation is the use of the graphical representation of the basis to carry out the revised simplex operations. This is a means of exploiting the special structure and sparseness of the constraint matrix to minimize computational effort and storage requirements. We also present the results of solving several large problems with the code developed.

INTRODUCTION

The capacitated transportation problem deals with the optimal allocation of resource S_i available at source $i \epsilon N_S = \{1, \ldots, n_s\}$ to meet the demand T_j at destination $j \epsilon N_T = \{1, \ldots, n_t\}$. We assume that each source is connected by arcs to one or more destinations, and supply between sources or between destinations is not possible. We denote the set of arcs by P and the unit cost for arc a_k by c_k, $k \epsilon P$. We will also require that the flow x_k along arc a_k be nonnegative and does not exceed an upper bound M_k. We denote the set of arcs connected to source i by Q_i and the set of arcs connected to destination j by D_j. The problem can then be stated as:

(1) (T.P.) $$\text{Minimize} \sum_{k \epsilon P} c_k X_k$$

(2) $$\text{subject to:} \sum_{k \epsilon Q_i} x_k \leq S_i, \qquad i \epsilon N_S$$

$$\sum_{k \in D_j} - x_k = - T_j, \qquad j \in N_T \tag{3}$$

$$0 \leq x_i \leq M_i, \qquad i \in P. \tag{4}$$

There have been several recent studies, including the present one, which attempt to solve network problems by using the graphical representation of basic solutions. Dantzig [2] and Johnson [10] describe the representation, and some of the relevant terminology and results used in this study are summarized below.

In the graphical representation of T.P., the sets N_S and N_T define a finite set of *nodes* (vertices). The arcs in P join a node in N_S to a node in N_T and will be called *regular arcs*. Since each variable appears in one equation in (2) with a $+1$ coefficient, and in one equation in (3) with a -1 coefficient, all the regular arcs are *directed* from a node in N_S to a node in N_T. When a slack variable is added to each equation in (2), each node in N_S has a *slack arc* connected to it. Such slack arcs are connected to that node only. Note that no node in N_T has a slack arc connected to it. The set of nodes and arcs constitute a *graph*.

Now consider a graph with associated nodes and arcs. A *simple path* in a graph is a set of nodes and connecting arcs of the graph such that no node (or arc) is repeated. A simple *cycle* is a simple path together with an arc joining the beginning node and ending node of the path. A *connected graph* has at least one path between every pair of nodes, and a *tree* is a connected graph with no cycles. A *forest* is a graph consisting of one or more trees, and a *spanning forest* is a forest whose node set equals the node set of the graph. A tree with a slack arc associated with it is called a *rooted tree* and the slack arc is called the *root* of the tree.

Using the above definitions, we can identify an appropriate spanning forest with each basic solution of T.P. Let B denote the "working" *basis* of T.P. For a nondegenerate solution this consists of the columns in (2) and (3), including slack columns, associated with $0 < x_j < M_j$. A subgraph F_B associated with B is composed of the arcs corresponding to the columns of B and the nodes on which these arcs are incident. Johnson [10] has proved certain important properties of the subgraph F_B, namely

(i) If B is a basis, F_B is a spanning forest, and

(ii) Each tree of F_B contains one and only one basic slack arc (root)

This result can be used to carry out the simplex operations efficiently to solve T.P. The effectiveness arises from the fact that by dealing with the subgraph F_B, one essentially can exploit the sparseness of the constraint set (2) and (3). The graph can also be viewed as a means for identifying the sets of primal and dual variables whose values change at each simplex iteration. These sets, in general, will be much smaller than the complete set of primal variables x_i, $i \in P$, or $n_s + n_t$ dual variables.

DUAL VARIABLE OPERATIONS

Consider a basis forest F_B which consists of one or more basis trees. Let arc a_1 be the root of one of the trees with arc a_1 touching node n_1. Let n_p be an arbitrary node in the tree. Consider the unique path from n_1 to n_p, and denote the arcs in the path by $a_2, \ldots, a_p$.

To calculate the *simplex multipliers* associated with the basis B, we need to solve the equation $\Pi B = c_B$ which can be expressed by a suitable partitioning as

$$(\Pi_p | \Pi_q) \left[\frac{B_p}{0} \middle| B' \right] = (c_p | c_q), \tag{5}$$

where

$$B_p = \begin{array}{c} \\ 1 \\ \vdots \\ \\ p \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & \cdots & & & k & \cdots & p \end{array} \\ \begin{bmatrix} 1 & 1 & & & & & \\ -1 & & -1 & & & & \\ & \ddots & & \ddots & & & \\ & & & & (-1)^k & & \\ & & & & (-1)^{k+1} & & \\ & & & & & \ddots & \\ & & & & & & \end{bmatrix} \end{array}$$

correspond to the variables in the path, and B' denotes the remaining columns in the basis. The solution to (5) clearly must satisfy $\Pi_p B_p = c_p$. That is, we must have

(6a) $$\Pi_1 = c_1$$

(6b) $$\Pi_{k-1}(-1)^k + \Pi_k(-1)^{k+1} = c_k \qquad 2 \leq k \leq p.$$

By defining the *relative cost factors* $\bar{c}_k$ for an arc a_k as

(7) $$\bar{c}_k = \begin{cases} (c_k - \Pi_x) \text{ for slack arc } a_k \text{ touching node } x \epsilon N_S \\ (c_k - \Pi_x + \Pi_y) \text{ for regular arc } a_k \text{ touching nodes } x \epsilon N_S,\ y \epsilon N_T, \end{cases}$$

Equations (6) can be expressed more compactly as $\bar{c}_k = 0$, $k = 1, \ldots, p$. Note that, in our formulation $c_k = 0$ when a_k is a slack arc. Starting with Π_1, Equation (7) can be used to compute iteratively the simplex multipliers $\Pi_1, \ldots, \Pi_p$ for the nodes in the path. Since a simple path exists for each node in the basis forest to some root, the above equations can also be used to calculate the initial simplex multipliers, and is the procedure given by Johnson [10].

In the bounded variable simplex method, a nonbasic variable is chosen to enter the basis which does not satisfy the optimality criteria:

(8) $$x_k = 0 \text{ implies } \bar{c}_k \geq 0 \text{ and } x_k = M_k \text{ implies } \bar{c}_k \leq 0,$$

where $\bar{c}_k$ is given by (7). For a slack variable, we cannot have $x_k = M_k$ and only the first criterion applies.

Given a basis B, suppose arc a_J (variable x_J) in the basis forest F_B is the leaving arc. A node n_i is said to be *above* arc a_J (with respect to the basis B) if the unique simple path from n_i to the root (of the corresponding tree) contains a_J. We will denote the set of all such nodes by N', i.e.,

(9) $$N' = \{i: \text{ node } n_i \text{ is } \textit{above} \text{ arc } a_J \textit{ before} \text{ the basis change}\}.$$

Clearly, if a_J is a basic slack arc, all the nodes in the tree containing a_J are above a_J. Proposition 1 below specifies which simplex multipliers Π_i change when some variable x_E replaces variable x_J in the basis. The corollary gives the corresponding changes in the relative cost factors.

PROPOSITION 1: Let a variable x_J in the basis be replaced by a variable x_E with relative cost

factor $\bar{c}_E$ defined by (7). Suppose a_E touches node $n_l \epsilon N_S$. Let the simplex multipliers before and after the basis change be denoted by Π_i and $\hat{\Pi}_i$. Then

$$\hat{\Pi}_i = \begin{cases} \Pi_i & i \notin N' \\ \Pi_i + \bar{c}_E & \text{if } i \epsilon N' \text{ and } n_l \epsilon N' \\ \Pi_i - \bar{c}_E & \text{if } i \epsilon N' \text{ and } n_l \notin N'. \end{cases} \tag{10}$$

PROOF: Consider a node n_i and the path from n_i to the root after the basis change. If $i \notin N'$, this path is the same as before the basis change and does not contain the arc a_E. Hence from the discussion earlier (see Eq. (6)). we have

$$\hat{\Pi}_i = \Pi_i, \qquad i \notin N'. \tag{11}$$

Suppose a_E is a regular arc. Denote by $n_r \epsilon N_T$ the second node it touches. From Eq. (7), we have

$$\hat{\Pi}_{n_l} - \hat{\Pi}_{n_r} = c_E.$$

Subtracting $(\Pi_{n_l} - \Pi_{n_r})$ from both sides.

$$(\hat{\Pi}_{n_l} - \Pi_{n_l}) - (\hat{\Pi}_{n_r} - \Pi_{n_r}) = c_E - \Pi_{n_l} + \Pi_{n_r} = \bar{c}_E. \tag{12}$$

If $n_l \epsilon N'$, then $n_r \notin N'$, and $\hat{\Pi}_{n_r} = \Pi_{n_r}$ from (11), and (12) gives $\hat{\Pi}_{n_l} = \Pi_{n_l} + \bar{c}_E$. If $n_l \notin N'$, again from (11), $\hat{\Pi}_{n_l} = \Pi_{n_l}$ and from (12), we get $\hat{\Pi}_{n_r} = \Pi_{n_r} - \bar{c}_E$. Application of (7) recursively along the path away from the root then gives

$$\hat{\Pi}_i = \begin{cases} \Pi_i + \bar{c}_E & \text{for } i \epsilon N' \text{ and } n_l \epsilon N' \\ \Pi_i - \bar{c}_E & \text{for } i \epsilon N' \text{ and } n_l \notin N'. \end{cases} \tag{13}$$

If a_E is a slack arc, from (7), we see that (12) holds with Π_{n_r} and $\hat{\Pi}_{n_r}$ deleted. Hence, $\hat{\Pi}_{n_l} = \Pi_{n_l} + \bar{c}_E$ which is the same expression as that for the regular arc since $n_l \epsilon N'$ always for a slack arc. This completes the proof.

COROLLARY: Let S be the set of regular or slack arcs touching exactly one node in N'. Let the relative cost factors for an arbitrary arc a_i touching node $x \epsilon N_S$ be $\bar{c}_i$ and $\hat{\bar{c}}_i$, respectively, before and after the basis change. Then

$$\hat{\bar{c}}_i = \begin{array}{ll} \bar{c}_i & a_i \notin S \\ \bar{c}_i - \bar{c}_E & a_i \epsilon S,\ n_l,\ x \epsilon N' \text{ or } n_l,\ x \notin N' \\ \bar{c}_i + \bar{c}_E & a_i \epsilon S,\ n_l \notin N',\ x \epsilon N' \text{ or } n_l \epsilon N',\ x \notin N'. \end{array}$$

PROOF: From (7), we have for a regular arc joining nodes $x \epsilon N_S$ and $y \epsilon N_T$,

$$\hat{\bar{c}}_i = c_i - \hat{\Pi}_x + \hat{\Pi}_y = \bar{c}_i - (\hat{\Pi}_x - \Pi_x) + (\hat{\Pi}_y - \Pi_y),$$

and for a slack arc touching node $x \epsilon N_S$,

$$\hat{\bar{c}}_i = c_i - \hat{\Pi}_x = \bar{c}_i - (\hat{\Pi}_x - \Pi_x).$$

When $a_i \notin S$, we have $x \notin N'$ for a slack arc, and $x, y \epsilon N'$ or $x, y \notin N'$ for a regular arc. When $a_i \epsilon S$, we have $x \epsilon N'$ for a slack arc. In this case, for a regular arc $x \epsilon N'$ implies $y \notin N'$ and vice versa. Substituting the values of $(\hat{\Pi}_x - \Pi_x)$ and $(\hat{\Pi}_y - \Pi_y)$ from (10) for the various cases the result is immediate.

Glover, Klingman, and Karney [7] present the above results in a slightly different form for the uncapacitated transportation problem.* Proposition 1 and the corollary provide the means of updating the relative cost factors in two related ways. Both approaches use the graphical representation of the basis to identify the changes caused by a basis change. In our computational procedure we use Proposition 1 to recompute only the simplex multipliers that change.

PRIMAL VARIABLE OPERATIONS

The simplex multipliers discussed above and the relative cost factors provide the means of identifying nonoptimal basic solutions and the variable x_E to enter the basis. To determine the variable to leave the basis, we need to generate the current column of x_E. We discuss below how this can be done using the graphical representation of the basis.

Let x_E be a nonbasic variable incident on nodes $n_l \epsilon N_S$ and $n_r \epsilon N_{\bar{T}}$. Let P'_L be the set of basic variables corresponding to the unique path from n_l to the root including the basic slack arc at the root. Let P_R be similarly defined. Let

$$P_L = P'_L - (P'_L \cap P'_R) \quad \text{and} \quad P_R = P'_R - (P'_L \cap P'_R). \tag{14}$$

If $P'_L \cap P'_R \neq \phi$, then n_l and n_r are contained in the same tree. In this case, P_L and P_R corresponds to paths from n_l and n_r to a certain node n_t where these join. If $P'_L \cap P'_R = \phi$, then n_l and n_r are in different trees, and (14) gives $P_L = P'_L$ and $P_R = P'_R$. Without loss of generality, let us denote the nodes in path P_L by $n_1, \ldots, n_p$ and those in path P_R by $n_{p+1}, \ldots, n_{p+q}$. Note that since x_E is incident on nodes n_l and n_r, we have $n_l = n_1$ and $n_r = n_{p+1}$.

The above notation readily extends to the case where x_E is a nonbasic slack variable touching only a node $n_l \epsilon N_S$. In this case $P'_R = \phi$ and (14) gives $P_L = P'_L$ and $P_R = \phi$.

We will now state a proposition which gives a means of generating the updated column of a nonbasic variable x_E.

PROPOSITION 2: The original column k_E for a nonbasic variable x_E can be uniquely expressed as a linear combination of the basic variables $x_i \epsilon P_L \cup P_R$, i.e., the only nonzero entries in the current column for x_E are in the rows of $x_i \epsilon P_L \cup P_R$. Further, each such row contains a nonzero entry in the current column of x_E.

PROOF: We will derive the unique solution to the equation

$$Bd = k_E. \tag{15}$$

*This equivalence can easily be established by a suitable redefinition of $c(q_1, q_2)$ in Corollary 3 of [7].

Without loss of generality, let the variables associated with P_L be $x_1, \ldots, x_p$, and those with P_R be $x_{p+1}, \ldots, x_{p+q}$ ordered from n_l and n_r toward the root(s). Rearrange the columns and rows of B in the order $x_1, \ldots, x_p, x_{p+1}, \ldots, x_{p+q}, \ldots$, and $n_1, \ldots, n_p, n_{p+1}, \ldots, n_{p+q}, n_t, \ldots$ respectively where n_t is the node where paths P_L and P_R join when $P'_L \cap P'_R \neq \phi$. Then (15) can be expressed as*:

$$
\begin{array}{c}
\qquad 1 \ldots p \quad p+1 \ldots p+q \quad \ldots \\
\begin{array}{c} 1 \\ \vdots \\ p \\ p+1 \\ \vdots \\ p+q \\ t \\ \vdots \end{array}
\left[\begin{array}{c|c|c}
B_L & 0 & \\
\hline
0 & B_R & B' \\
\hline
D & &
\end{array}\right]
\left(\begin{array}{c} d_L \\ \hline d_R \\ \hline d' \end{array}\right)
\left(\begin{array}{c} 1 \\ 0 \\ \vdots \\ 0 \\ \hline -1 \\ 0 \\ \vdots \\ 0 \\ \hline 0 \end{array}\right)
\end{array}
\tag{16}
$$

where B_L and B_R have the form:

$$
B_L = \begin{array}{c} 1 \\ \\ \\ \\ \\ \\ p \end{array}
\left[\begin{array}{cccccc}
1 & & & & & \\
-1 & -1 & & & & \\
 & 1 & 1 & & & \\
 & & \ddots & \ddots & & \\
 & & & & (-1)^{p-2} & \\
 & & & & (-1)^{p-1} & (-1)^{p-1}
\end{array}\right]
\tag{17}
$$

and

$$
B_R = \begin{array}{c} p+1 \\ \\ \\ \\ \\ \\ p+q \end{array}
\left[\begin{array}{cccccc}
-1 & & & & & \\
1 & 1 & & & & \\
 & -1 & -1 & & & \\
 & & \ddots & \ddots & & \\
 & & & & (-1)^{q-1} & \\
 & & & & (-1)^{q} & (-1)^{q}
\end{array}\right]
\tag{18}
$$

* Note that when x_E is a slack variable $P'_R = P_R = \phi$ and hence $B_R = \phi$, and the right-hand side of (16) has only one nonzero entry, a one in the first row.

The matrix D has all zero entries when $P'_L \cap P'_R = \phi$. When $P'_L \cap P'_R \neq \phi$, the matrix has zero entries except in the row of n_t. In this row, the nonzero entries are $(-1)^p$ and $(-1)^{q+1}$ in columns of x_p and x_{p+q}.

To solve (16), we first set $d' = 0$. When $P'_L \cap P'_R = \phi$, D is a zero matrix and the solution is obtained by solving

(19a) $$B_L d_L = \begin{pmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix}$$

and

(19b) $$B_R d_R = \begin{pmatrix} -1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix}$$

When $P'_L \cap P'_R \neq \phi$, the equation from row corresponding to n_t yields:

$$(-1)^p d_p + (-1)^{q+1} d_{p+q} = 0.$$

But this is satisfied by the solution of (19) since adding the equations in (19) gives

$$(-1)^{p-1} d_p + (-1)^q d_{p+q} = 0.$$

Hence, we only need to solve Eq. (19). From (17) and (19a), the solution is

(20) $$d_1^* = 1 \text{ and } d_i^* = -d_{i-1} \qquad i = 2, \ldots, p.$$

Again, from (18) and (19b), we get

(21) $$d_{p+1}^* = -1 \text{ and } d_{p+i}^* = -d_{p+i-1} \qquad i = 1, \ldots, q.$$

Thus d_i^* given by Eqs. (20) and (21) and $d' = 0$ solves Eq. (16) in all cases. But since B is a basis for T.P., this solution is unique and is equal to $B^{-1}k_E$ and (20) and (21) give the nonzero entries for the current column of x_E in terms of variables in P_L and P_R.

It may be noted that the above proof readily yields the method of generating a column of B^{-1}. In this case, instead of Eq. (15), we seek the solution to

$$Bd = (0, \ldots, 1, \ldots 0)'.$$

When a nonbasic variable x_E enters the basis, Eq. (20) and (21) above give the nonzero elements of the current column of x_E. These elements are in the paths P_L and P_R defined earlier. First consider the case when $x_E = 0$, and is to be increased. One of the variables in this path will either reduce to zero

or reach its upper bound. This particular variable will then be the leaving variable for the simplex method.

To identify the leaving variable, let $R^+ = \{i : d_i^* = 1, i = 1, \ldots, p+q\}$ and let

$$R^- = \{i : d_i^* = -1, i = 1, \ldots, p+q\}.$$

The x_i, $i \epsilon R^+$ will decrease when x_E is increased and x_i, $i \epsilon R^-$ will increase when x_E is increased. The maximum amount x_E can be increased is given by:

$$\Delta = \text{Min} \left\{ \underset{i \epsilon R^+}{\text{Min}} (\bar{x}_i), \underset{i \epsilon R^-}{\text{Min}} (M_i - \bar{x}_i), M_E \right\}, \tag{22}$$

where $\bar{x}_i$ are the current values of the variables. If M_E is the minimum in (22), the new basic solution is obtaining by adding Δ to x_i, $i \epsilon R^-$ and subtracting Δ from x_i, $i \epsilon R^+$. Since no basis change is involved, the simplex multipliers do not change. Otherwise, the element $J \epsilon R^+ \cup R^-$ where the minimum in (22) is achieved, identifies the leaving variable x_J. A basis change is required, and the simplex multipliers can be changed according to Proposition 1.

When $x_E = M_E$, x_E is to be decreased and (22) becomes

$$\Delta = \text{Min} \left\{ \underset{i \epsilon R^+}{\text{Min}} (M_i - \bar{x}_i), \underset{i \epsilon R^-}{\text{Min}} (\bar{x}_i), M_E \right\} \tag{23}$$

and changes in the simplex multipliers can be affected as before.

PRIMAL ALGORITHM

The dual and primal operations discussed earlier rely heavily on the graphical characterization of the basis. These can be used to implement the revised simplex algorithm using the graphical representation of the problem. Assume that an initial feasible basis and the associated values of the primal variables x_i and the simplex multipliers $\prod_i$ are known. The details of the algorithm are as follows:

1. Determine $\{\bar{c}_i\}$ by (7). If all $\bar{c}_i$ satisfy (8), the solution is optimal. Otherwise, select a x_E such that (8) is violated. If x_E is a slack variable go to Step 4; otherwise go to Step 2.

2. Trace paths P_L' and P_R' from both ends n_l and n_r of x_E to a root. Let $P_L \leftarrow (P_L' - (P_L' \cap P_R'))$ and $P_R \leftarrow (P_R' - (P_L' \cap P_R'))$. Go to Step 3.

3. Determine Δ by (22) or (23). Update the values for $\bar{x}_i$ for a_E and the arcs in paths P_L and P_R. If variable x_J is replaced by x_E, update the basis representation (see next section) and change the simplex multipliers for nodes $i \epsilon N'$ by (10). Return to Step 1.

4. If x_E touches n_l, determine the path P_L' from n_l to the root. Let $P_L \leftarrow P_L'$ and $P_R \leftarrow \phi$ and go to Step 3.

Basis Representation

The basis is maintained in graph form using the triple label scheme of Johnson [11]. The method requires three labels for each node (equation): a down label D_i, an up label U_i, and a right label R_i. The down label D_i is the unique node below node i in a tree. If D_i is zero, node i has a basic slack arc incident on it and is the root node. Since more than one node may be above a node i in a tree the

specific up and right labels are not unique. In general U_i is a node directly (connected by an arc) above node i in a tree. Label R_i indicates that node R_i is above the same node D_i as node i. A downward trace is accomplished using the down labels only. The method of tracing all nodes above the specified node is given in Johnson [11]. The updating of labels is straightforward and details are given by Glover and Klingman [4] and Srinivasan and Thompson [13]. It should be mentioned that the only labels that must be updated for a basis change are those representing the hierarchy between the entering and leaving variables.

In our implementation, the basic labeling system of Johnson was modified slightly. Given that two nodes are adjacent in a tree, the particular arc between them must be specified. For instance in determining the leaving variable one needs the value of the variable corresponding to the arc joining node i and node D_i. For this purpose we added an additional label A_i to specify the arc between nodes i and D_i. The updating of this label is again straightforward and accomplished in the same manner as that for the down labels.

STORAGE REQUIREMENTS

For the uncapacitated version of the code, a problem with m nodes and n arcs requires one list of length n to store the costs c_i. Further 10 lists of length m are needed as follows: (a) Four for recording the basis, (i.e., three for down, up and right nodes, and one for the down arc), (b) Two for variables values, basic primal and dual (simplex multipliers), and (c) Four temporary lists for column construction. Thus for 100-percent dense problems, i.e., if all the n_s nodes in N_S are connected with all the n_t nodes in N_T, the storage requirement is $10n_s + 10n_t + n_sn_t$.

For the capacitated version two additional lists of length n are required to indicate the bounds and whether a variable is basic or nonbasic at its upper or lower bound. For less than 100-percent dense problems, two additional lists, one of length $n_s(n_t)$ and one of length n can be used to allow the explicit consideration of only the n variables which are present. The $n_s(n_t)$ length list will indicate the number of variables in each source (destination) equation. The n length list will indicate the destination (source) equation for each nonslack variable. Then knowing which source x and destination y a nonbasic arc a_i connects, the relative cost factor can be calculated and an entering variable selected.

COMPUTATIONAL EXPERIENCE

Alternative means of obtaining an initial feasible basis and alternatives for selecting the variable to enter the basis are presented in the literature, see [3, 6, 9, 14]. Glover, et al. [6] and Srinivasan and Thompson indicated that the *Mofified Row Minimum Scheme* (MRM) [14] for determining a starting basis and the *Modified Row First Negative Evaluator Rule* (MRFN) [3] for selecting the pivoting variable are superior to the other rules that have been proposed. However, the results in Table 1 indicate that

TABLE 1. *Comparison of solution times as a function of oversupply*

Size	$0 \leq F \leq 0.1$		$0.1 < F \leq 0.2$		$0.2 < F \leq 0.5$	
	*TP*1	*TP*2	*TP*1	*TP*2	*TP*1	*TP*2
100 × 100	5528	4440	6449	6768	4012	4342
120 × 120	6940	6609	5814	6961	5986	5516
140 × 140	8976	7988	8820	8997	4733	4667
160 × 160	13386	11050	12638	12502	7685	8629

starting with artificial variables at all destinations and using the big M method coupled with the MRFN evaluator rule is comparable to the MRM start with the MRFN evaluator rule.

This study reports the computational experience with two codes for solving T.P. written in Fortran for the Univac 1108. *TP*1 uses the modified row minimum start rule and the modified row first negative evaluator choice rule. *TP*2 uses the big M method and the same choice rule. The object programs and data are held in fast speed core with all data in fixed point mode. Both programs are designed to solve problems whose product, mn, is less than or equal to 25,600. That is, the largest square problem which can be solved is 160×160. The total number of 36 bit words required for each program with data is approximately 62K. In addition to problem size, the effect of the extent of oversupply was also investigated. Oversupply can be expressed by

$$F = \left(\sum_i S_i - \sum_j T_j \right) \Big/ \sum_i S_i.$$

The problems generated were in one of the following categories (a) *tightly constrained*, $0 \leq F \leq 0.1$, (b) *moderately constrained*, $0.1 < F \leq 0.2$, or (c) *loosely constrained*, $0.2 < F \leq 0.5$. The supply for tightly and moderately constrained problems was uniformly distributed over the interval 10 to 50. The supply of the loosely constrained problems varied uniformly over the interval 30 to 70. All demands and costs were uniformly distributed over the interval 10 to 50. All problems which had no feasible solution were discarded.

Table 2 presents a summary of computational times as a function of problem size. Each of the averages are for 15 problems: five tightly constrained, five moderately constrained, and five loosely constrained. The times are in milliseconds and exclude input/output, but include all set-up.

TABLE 2. *Consolidated computational results*

Size	*TP*1		*TP*2	
	Average number pivots	Average time	Average number pivots	Average time
100×100	325	5330	446	5110
120×120	318	6247	482	6362
140×140	400	7509	601	7051
160×160	467	11236	653	10760

Table 1 presents a break-down of the results in Table 2 to highlight the effect of oversupply. It is interesting to note that *TP*2 solved tightly constrained problems somewhat faster than moderately constrained problems. As might be expected, the loosely constrained problems were easiest to solve.

In the above procedure we have represented the basis graphically and the means of identifying the variable to enter the basis is readily available. The nonzero elements in the column of this variable is recognized by tracing certain paths which yields the leaving variable. The updating of the basis representation again is easily accomplished. Only the simplex multipliers that change are recalculated at this stage. This results in a procedure that is computationally more efficient than other methods. The procedure can be extended to more general problems such as the minimum cost flow problem and the flow-with-gains problem, and is the subject of a forthcoming paper.

REFERENCES

[1] Charnes, A. and W. W. Cooper, "The Stepping-Stone Method of Explaining Linear Programming Calculations in Transportation Problems," Management Science *1*, 49–69 (1954).

[2] Dantzig, G., *Linear Programming and Extensions* (Princeton University Press, 1963).

[3] Dennis, J. B., "A High-Speed Computer Technique for the Transportation Problem," J. Assoc. Computing Machinery *8*, 132–153 (1958).

[4] Glover, F. and D. Klingman, "Locating Stepping-Stone Paths in Distribution Problems Via the Predecessor Index Method," Transportation Science *4*, 220–225 (1970).

[5] Glover, F., D. Klingman, A. Napier, and D. Karney, "A Comparison of Computation Times for Various Starting Procedures, Basis Change Criteria, and Solution Algorithms for Distribution Problems," Report CS 44, Center for Cybernetic Studies, University of Texas—Austin (1971).

[6] Glover, F., D. Karney, D. Klingman, and A. Napier, "A Computation Study on Start Procedures, Basis Change Criteria, and Solution Algorithms for Transportation Problems," Report CS 93, Center for Cybernetic Studies, University of Texas—Austin (1972).

[7] Glover, F., D. Klingman, and D. Karney, "The Augmented Predecessor Index Method for Locating Stepping-Stone Paths and Assigning Dual Prices in Distribution Problems," Transportation Science *6*, 171–179 (1972).

[8] Glover, F., D. Klingman, and D. Karney, "Double-Pricing Dual and Feasible Start Algorithms for the Capacitated Transportation (Distribution) Problem," Report CS 105, Center for Cybernetic Studies, University of Texas—Austin (1970 Rev. Oct. 1972).

[9] Hadley, G., *Linear Programming* (Addison-Wesley Publishing Company, Inc., Reading, Mass., 1962).

[10] Johnson, Ellis, "Programming in Networks and Graphs," ORC Report 65–1, University of California—Berkeley (1965).

[11] Johnson, Ellis, "Networks and Basic Solutions," Operations Research *14*, 619–623 (1966).

[12] Klingman, D., A. Napier, and J. Stutz, "NETGEN—A Program for Generating Large Scale (Un)Capacitated Assignment, Transportation and Minimum Cost Flow Network Problems," Report CS 109, Center for Cybernetic Studies, University of Texas—Austin (1973).

[13] Srinivasan, V. and G. L. Thompson, "Accelerated Algorithms for Labeling and Relabeling of Trees, with Applications to Distribution Problems," J. ACM *19*, 712–726 (1972).

[14] Srinivasan, V. and G. L. Thompson, "Benefit-Cost Analysis of Coding Techniques for the Primal Transportation Algorithms," J. ACM *20*, 194–213 (1973).

THE ALLOCATION OF RECRUITERS AMONG SPATIAL AREAS

James T. Bennett and Sheldon E. Haber

Program in Logistics

The George Washington University

1. INTRODUCTION

In their paper on the economics of the military draft [4], Hansen and Weisbrod discuss a number of inefficiencies of the military draft system. The principal inefficiency noted is the loss in output resulting from the utilization of labor services of draftees whose productivity in the civilian sector is, on the average, higher than that of individuals who would volunteer for military service.

With the advent of the all-volunteer military establishment, attention has shifted to the expenditures that will be necessary to meet military manpower requirements. A major problem to be resolved is the allocation of expenditures among competing uses, e.g., military pay, advertising and recruiting. Analysis of this problem is, of course, enormously difficult.* Because the problem is complex, only one aspect is considered here, that of suboptimization of a given resource, namely, the labor services of recruiters. The particular aspect studied is the efficient allocation of recruiters among geographic areas. The importance of analyzing this facet of military manpower procurement is heightened by recent decisions of the Department of Defense to increase the number of recruiters assigned to each of the services.†

To a greater or lesser extent, depending on the branch of service, the current procedure for distributing recruiters is to allocate them to recruiting markets in proportion to market size as measured by the number of "qualified military availables" (QMA) in an area.* This rule is consistent with efficient utilization of recruiters when all areas are alike in terms of ease of recruitment. If, however, it is easier to recruit volunteers in some areas than in others, the application of the rule results in differences in recruiter productivity among areas. One important consequence is that some individuals will be induced to enlist from areas where their productivity in the civilian sector is high, while other individuals, whose contribution to output is less and who therefore might be more disposed to enlisting, would fail to be recruited because of an insufficiency of recruiters in their area. The resultant loss in output is similar to that described in [4]. Another important consequence is that the number or quality

*To date, emphasis has been given to the level of military pay needed to raise an all-volunteer military service, as for example in [1, 2, 3]. The relationship between enlistments and other types of expenditures has received little attention.

†In fiscal year 1971, 7,500 servicemen were employed as recruiters. Although the Armed Forces is probably the largest employer of recruiters, the problem of spatial allocation of recruiters is of importance to other organizations as well, e.g., to universities who recruit students on a regional or national basis, and is a special case of the more general problem of spatial allocation of factor inputs.

*The QMA for an area is the estimated number of males age 17–21 who can pass both the mental and physical tests for military service.

of enlistments is reduced when recruiter allocation is not economically efficient: both can be raised by redistributing recruiters to market areas where the productivity of recruiters is high.

In the next section, similarities among the services in the allocation of recruiters to recruiting areas in terms of market size are documented. In section 3, it is shown that the current method of allocating recruiters leads to wide variations among areas in recruiter productivity. Some theoretical aspects of the optimal distribution of recruiters are discussed in section 4. An important theoretical contribution made in this section is that the conventional textbook rule for maximizing output, by allocating recruiters until marginal output is equal in each spatial area, is invalid if recruits are terminated from service because, for example, they are unable to complete basic training. In this case, which is not uncommon, the problem of optimal allocation is not straightforward conceptually or empirically. In section 5, summary comments are presented.

2. THE DISTRIBUTION OF RECRUITERS BY GEOGRAPHIC AREA

In the past, the basis for the geographical allocation of recruiters has been the number of QMA's in an area. The rationale for using QMA data to allocate recruiters is that it provides a measure of market size. Market size, however, is not equivalent to market potential, and it is market potential rather than size which is of importance in maximizing recruiter efficiency.

Market potential is in part related to the number of candidates for enlistment, but it is also determined by the propensity of those candidates to enlist in the armed forces. To the degree that the distribution of recruiters is proportional to QMA in an area and enlistment rates differ among areas, recruiter productivity will vary among recruiting markets. Efficient allocation requires that recruiters be redistributed from areas where the cost of recruitment is high to areas where it is low, and therefore the current procedure leads to inefficient utilization of recruiters.

That recruiters are allocated in proportion to an area's QMA can be seen from Table 1 which shows the number of recruiters per 1,000 QMA for the recruiting markets of each service.* For each of the services there is a marked concentration in the ratio within a narrow range. For example, in 55.6 percent of the Air Force recruting areas the ratio was between 0.21 and 0.30. One notices, however, variations in the ratio. In remote and sparsely populated areas where large distances must be covered to contact prospective enlistees, the cost of recruitment is high; as a consequence the ratio of recruiters to QMA may be low. Conversely, in areas where the density of population is low, the ratio of recruiters to QMA may be high. While there are differences in the ratio of recruiters to QMA, the narrow range of concentration in this ratio is to be emphasized.

3. VARIATIONS IN RECRUITER PRODUCTIVITY BY GEOGRAPHIC AREA

In determining the extent to which recruiter productivity varies among areas, several problems must be faced including the problem of defining inputs and outputs. The usual measure of labor input

*The data for recruiters are reported for 46 Air Force, 40 Army, 41 Marine, and 37 Navy recruiting markets. These range in size from areas that are smaller than a state, as in the Northeast, to areas that encompass several states, as in the Northwest. The geographical boundaries of these markets differ. For many areas, market boundaries are similar among the services; for others there is little correspondence. In general, commonality of market areas is greatest for the Army, Marines, and Navy, and least for the Air Force. The data for QMA, on the other hand, are reported by state. For this reason, it is necessary to aggregate recruiting markets into observational units comprising one or more states. Because of the differences in boundaries of recruiting markets, the number of observations differ for each service. The number of observations are 27, 29, 33, and 29, for the Air Force, Army, Marines, and Navy, respectively.

TABLE 1. *Percentage frequency distribution of recruiting areas, calendar year 1970*

Recruiters[a] per 1,000 QMA	Percentage distribution of recruiting areas			
	Air Force	Army	Marines	Navy
0.10 or less	—	—	—	—
0.11–0.20	3.7	—	15.2	—
0.21–0.30	55.6	—	51.5	13.8
0.31–0.40	29.6	10.3	21.2	13.8
0.41–0.50	7.4	55.2	9.1	37.9
0.51–0.60	3.7	31.0	3.0	10.3
0.61–0.70	—	3.5	—	13.8
0.71–0.80	—	—	—	6.9
0.81–0.90	—	—	—	3.5
Total	100.0	100.0	100.0	100.0

[a] For each service except the Air Force, the number of recruiters is an average for fiscal years 1970 and 1971. For the Air Force, only fiscal year 1971 figures are used.

is in terms of man-hours. Data for this measure are not available, however, due to a lack of man-hour accounting of hours worked by recruiters. In the absence of such information, labor input can only be measured by counting recruiters. With respect to output, alternative measures are possible. The most common measure is the number of enlistees recruited per recruiter, which is a measure of the gross productivity of recruiters. Productivity, however, can also be measured in net terms. Implicit to this latter concept is the notion that the labor services of enlistees become effective only after demonstration of some minimum level of capability.*

Another problem is that of the quality of inputs and outputs. Although it is reasonable to assume that geographical differences in recruiter quality are not important, a similar assumption with respect to quality of output can only be made with some degree of risk. For example, it is not unlikely that the quality of labor services provided by enlistees is related to level of education,† and there is abundant evidence that educational attainment varies by geographic area. Nonetheless, because of the difficulty in measuring the quality of recruits, it is assumed that it is uniform in all market areas.**

Geographic differences in the gross productivity of recruiters by service are shown in Table 2. A number of conclusions can be drawn from this table. First, regional differences in productivity (Col. 1) varied among the services;†† they were smallest for the Navy and largest for the Army. Second,

*One common practice used in defining net productivity is to measure the output of recruiters in terms of the number of recruits they enlist who passed basic training. Other definitions of output are also possible, e.g., the number of enlistees who fullfill their enlistment contract, but are not as timely for management decision making.

†See [5], pp. 77 and 84–86.

**The problem of quality is less acute when productivity is defined in net terms, e.g., net of individuals who subsequently fail to pass training requirements, instead of gross terms where each recruit represents a unit of output, since as noted the net measure implies a minimum acceptable level of quality.

††As with individual recruiting markets, the geographical boundaries of regional recruiting districts differ by service. These differences are best illustrated by noting that the number of regional districts ranges from five for the Army to eight for the Navy with no two services having the same number of districts. To facilitate comparisons among the services, the gross productivity figures are shown for the four major regions distinguished in the Census of Population.

the generally lower gross productivities of recruiters in the Northeast and generally higher productivities in the South† suggest that a redistribution of recruiters from the Northeast to the South can lead to increased enlistments. Third, differences in gross productivity within regions were larger than the differences among regions.** For example, for the Marines, gross productivity in the Northwest ranged from 34.7 enlistees per recruiter in the recruiting area where productivity was lowest to 68.4 in the area where productivity was highest. Thus, while one recruiter recruited approximately 70 enlistees in the most productive recruiting area, two recruiters were needed to recruit the same number in the least productive area. From these data, the important conclusion is drawn that the efficiency of recruitment can be increased in each of the services by redistributing recruiters among geographical areas.

TABLE 2. *Gross productivity of recruiters, calendar year 1970*

Region and Service	Average	Minimum	Maximum
Northeast:			
Air Force [a]	45.8	41.2	56.6
Army	46.3	42.6	66.7
Marines	35.5	30.1	38.9
Navy	39.0	37.1	40.1
Northwest:			
Air Force	49.2	45.8	53.9
Army	61.1	49.0	66.6
Marines	45.8	34.7	68.4
Navy	37.4	28.9	54.9
South:			
Air Force	54.4	31.4	71.2
Army [b]	62.7	52.2	78.5
Marines	47.4	33.0	60.4
Navy [b]	39.2	27.0	56.2
Far West:			
Air Force	46.8	34.0	63.3
Army	58.8	48.5	72.4
Marines	36.3	24.3	40.0
Navy	35.9	32.5	40.3

[a] See footnote a, Table 1.
[b] Includes New Mexico.

Implicit to the above discussion are two assumptions. First, that the variations in gross productivity reflect differences among areas in the propensity of individuals to enlist in the armed forces. Second, that the redistribution of recruiters from market areas where gross productivity is low to areas where it is high leads to a reduction in recruiter productivity in the latter and an increase in recruiter pro-

† It is interesting to note that outside the South, the gross productivity of recruiters was highest in the Northwest for three of the four services.

** This is indicated by the last two columns in Table 2 where the minimum and maximum gross productivity of recruiting areas are shown for each region.

ductivity in the former. These suggest that given the ratio of recruiters to QMA, the gross productivity of recruiters will vary positively with the propensity to enlist in an area; and given the propensity to enlist and that the services are operating in the diminishing returns portion of their production function, gross productivity will vary inversely with the concentration of recruiters in an area. The data in Table 3 confirm these expectations. In this table, gross productivity is shown as a function of the number of recruiters per 1,000 QMA and the enlistment rate* as given in the table stubs. For example, in Army recruiting areas with low ratios of recruiters to 1,000 QMA (less than 0.480) and low enlistment rates (less than 27.9), gross productivity was 59.8. With only one exception, gross productivity rose as the enlistment rate increased, and in all cases gross productivity fell as the concentration of recruiters increased.

As shown above, when recruiters are allocated spatially on the basis of market size without taking into account differences among areas in the propensity to enlist, the result is wide disparities in the gross productivity of recruiters. The data presented also suggest that recruiting is subject to diminishing returns and, hence, that the application of marginal productivity analysis to the related problems of optimal allocation of labor inputs and maximization of output is particularly appropriate.

TABLE 3. *Gross productivity for areas with low and high ratios of recruiters to QMA and low and high enlistment rates, calendar year 1970*

	Low E [b] ($E < 14.7$)	High E ($E \geq 14.7$)
Air Force:		
Low R [a] ($R < 0.296$)	51.8	65.0
High R ($R \geq 0.296$)	42.6	46.7
	Low E ($E < 27.9$)	High E ($E \geq 27.9$)
Army:		
Low R ($R < 0.480$)	59.8	66.6
High R ($R \geq 0.480$)	46.0	59.0
	Low E ($E < 11.1$)	High E ($E \geq 11.1$)
Marines:		
Low R ($R < 0.265$)	44.2	56.5
High R ($R \geq 0.265$)	34.7	37.2
	Low E ($E < 0.44$)	High E ($E \geq 0.444$)
Navy:		
Low R ($R < 0.444$)	43.5	39.7
High R ($R \geq 0.444$)	33.7	35.0

[a] R denotes the ratio of recruiters to 1,000 QMA.
[b] E denotes the ratio of enlistments to 1,000 QMA.

*The enlistment rate is used as a proxy for the propensity to enlist in an area and is defined as the number of total enlistments per 1,000 QMA.

4. SOME THEORETICAL ASPECTS OF THE OPTIMAL DISTRIBUTION OF RECRUITERS

In the foregoing, it was indicated that the procedure of allocating recruiters in proportion to QMA is inefficient. If variations in marginal gross productivity, i.e., the number of enlistees that would be recruited by an additional recruiter, are as large as the variations in average gross productivity among and within regions, increased enlistments* can result if recruiters are redistributed from areas where productivity is low to areas where it is high. However, although this general rule appears to provide a criterion for optimal allocation of recruiters, the problem of optimal allocation is not an easy one. The difficulty is related to the difference between gross and net productivity and its implication for optimal distribution in a situation where the market value of the services provided by enlistees cannot be easily estimated.

In the simple case where gross and net productivity are equal, optimal allocation occurs when marginal gross productivity (MGP) is the same in all market areas. If this condition is satisfied, output is maximized for a given level of recruiter resources. This is illustrated in Table 4 where enlistments are shown as a function of the number of recruiters in an area. For convenience, it is assumed that there are two market areas, each with the same QMA. Area I is an easy area to recruit in, Area II a difficult one. It is also assumed that there are a total of four recruiters to be allocated between the two areas. As can be seen from Table 4, when the number of recruiters is increased from 2 to 3 in Area I, total enlistments increase from 700 to 950, and marginal gross productivity, the difference in total enlistments, equals 250. From Table 5, one notes that MGP is 250 in both areas when three recruiters are assigned to Area I and one recruiter is assigned to Area II. For the given number of recruiters, four, total enlistments are a maximum, 1200, when MGP is the same in both areas. Any other distribution of recruiters, e.g., two in Area I and two in Area II, which would be the distribution if recruiters were allocated in proportion to QMA, would result in a smaller number of enlistments. Thus, although marginal and average productivity fall in Area I when the number of recruiters is increased from 2 to 3, marginal and average productivity rise in Area II, with the net result being an increase in total enlistments from 1,150 to 1,200.

When net productivity differs from gross productivity it might be thought that the same criterion for maximizing output applies, namely, that enlistments are maximized when marginal net productivity (MNP) is the same in all areas. But this is generally not the case as can be seen from the following illustration.

	Area I	*Area II*
(a) Marginal gross productivity.....	200	150
(b) Marginal net productivity	150	150

Although MNP is the same in both areas, MGP is higher in Area I indicating that the cost per effective enlistee (defined here as an enlistee who passes basic training) is higher in that area. In the example given, the cost of increasing armed forces strength by 150 men from Area II is the cost of maintaining one recruiter plus the training costs of 150 enlistees through basic training. The cost for Area I is larger

* Actually, for the services other than the Army which are almost wholly comprised of enlistees, further increased efficiency in recruiter allocation would not have resulted in a higher number of enlistments since their overall strength is fixed by Congress. Increased recruiter efficiency, however, would have permitted the recruitment of higher quality enlistees.

TABLE 4

Number of recruiters	Area I (QMA = 10,000)		Area II (QMA = 10,000)	
	Total enlistments	Marginal productivity	Total enlistments	Marginal productivity
0	0	—	0	—
1	400	400	250	250
2	700	300	450	200
3	950	250	600	150
4	1,100	150	700	100

TABLE 5

Recruiter combinations		Average productivity		Marginal productivity		Total enlistments, both areas
I	II	I	II	I	II	
0	4	0	175	—	100	850
1	3	400	200	400	150	1,000
2	2	350	225	300	200	1,150
3	1	317	250	250	250	1,200
4	0	275	0	150	—	1,100

than this amount by the cost of training an additional 50 enlistees for some portion of basic training.* Therefore, when net and gross productivity differ, the optimal allocation of recruiters depends on the difference between MNP and MGP as well as on MNP. As an example, net productivity data on Marine enlistees who passed basic training are presented in Table 6. Note that the ranking of districts on the basis of gross productivity, net productivity and the difference between the two measures of productivity is not the same. In particular, the Midwest ranks fourth on the basis of gross productivity, but second on the basis of net productivity. Also, although the net productivity measure is not too different for the Southeast and Far West, the difference between gross and net productivity is substantially smaller for the latter region. Thus, in the Southeast for every 43.2 enlistees who passed basic training, there were another 13.0 who did not finish basic training. In contrast, in the Far West only 6.4 enlistees failed basic training per 42.4 who passed.

In principle the values MNP and the difference between MGP and MNP can be transformed into a single measure, but in practice it cannot be done in an exact manner. For the case of output sold in the private market for which there is an established price, the cost of "defective" units can be subtracted from sales receipts in the same manner as any other cost of production to obtain a net revenue figure. Similarly, given the cost of training an enlistee of less than minimal quality, one can conceptually subtract this cost from the value of services rendered by enlistees to obtain an estimate of net benefits from such services, provided that there is an established market price which sets a

* To these costs should be added other costs such as the costs of initial processing, etc. Ideally, some adjustment should be made for the fact that the real cost of training enlistees who fail to pass basic training may be higher, per unit time period, than the costs of training the same number of enlistees who do pass basic training.

TABLE 6. *Gross and net productivity by Marine Corps District, fiscal year 1971*

Marine Corps District	Gross productivity [a]	Net productivity [b]	Gross-net productivity
Southwest	59.5	50.0	9.5
Southeast	56.2	43.2	13.0
Middle Atlantic	55.6	43.5	12.1
Midwest	55.0	47.0	8.0
Far West	48.8	42.4	6.4
Northeast	46.9	36.4	10.5

[a] Enlistees per recruiter.
[b] Enlistees who finished basic training per recruiter.

value on these services. Such a market price exists only for an all-volunteer military establishment. In this case, the value of services rendered by enlistees is the military wage paid them, which is the opportunity cost of employment in the civilian sector of the economy. The services rendered by an an enlistee, however, may not terminate after his first tour of duty. If he reenlists, these services can continue over a number of tours of duty. The exact present value of these future services is difficult to compute since *ceteris paribus* conditions do not prevail; e.g., on-the-job training increases the value of services rendered over time. Despite these problems, it is useful to draw the important distinction between gross and net productivity; to recognize that the equalization of net productivity does not, in general, yield the optimal distribution of recruiter resources; and to note that consideration must be given to the difference between gross and net productivity if economic efficiency is to be achieved.

5. SUMMARY

The main thrust of this paper has been to indicate the inefficiency inherent to the procedure of allocating recruiters among market areas in proportion to market size. This procedure assumes that market size and market potential are synonymous. They are not because measures of market size, such as qualified military availables (QMA) do not take into account variations among areas in the propensity to enlist. Moreover, the fact is ignored that relative changes in market size occur very slowly whereas relative changes in market potential can occur much more rapidly.

The procedure of allocating recruiters in proportion to QMA in an area is followed to a greater or lesser extent by all services. A principal consequence of this procedure is that it leads to marked differences in the gross productivity of recruiters due to differences in the propensity to enlist among market areas. These disparities in productivity suggest that the current procedure for spatially allocating recruiters is nonoptimal.

Although it is not very difficult to demonstrate that the market size criterion for allocating recruiters is inefficient, it is not a simple matter to formulate an exact procedure for achieving an optimal distribution of recruiters. The difficulty stems from the distinction between marginal gross productivity and marginal net productivity. When the two are equal, optimal allocation of recruiters is achieved when recruiters are allocated among areas such that gross productivity is equal in all areas. When gross and net productivity are not the same, which is the usual situation, the cost of enlistee "rejects" must be subtracted from the (gross) value of services provided by enlistees of minimal or higher quality to obtain a net value. These cost and value estimates are difficult to assess. If the objective of re-

cruiter allocations is efficiency,* the distinction between gross and net productivity must be recognized and taken into account. Failure to do so leads to a nonoptimal distribution of recruiters, and on the output side results in two important costs. First, the number of enlistees recruited or the quality of enlistees is less than what it would be if allocative efficiency were attained. Second, some individuals whose productivity in the civilian sector is high, who reside in areas where the recruiter to QMA ratio is unduly large, will be induced to enlist while other individuals, whose productivity in the civilian sector is low, will fail to be recruited because of an insufficiency of recruiters in their area.

In addition to the difficulties introduced by the need to consider net rather than gross productivity, other problems are encountered in allocating recruiters in an optimal manner. Not all potential enlistees are of equal quality, e.g., and enlistee's productivity in the military (as well as in the civilian) sector may depend on his level of education. The average length of enlistment is also likely to differ among areas, if only because length of enlistment is related to educational attainment. Moreover, just as there are differences among areas in the propensity to enlist, there may be differences in the propensity to reenlist. These qualitative aspects of force composition are of undeniable importance in meeting military manpower goals at reasonable cost, and complicate the problem of recruiter allocation since they suggest that there is more than one market for recruits. Finally, not only is the marginal productivity of recruiters influenced by the number of recruiters in an area, it is also influenced by such diverse factors as the military/civilian wage rate and the quality of management of recruiter resources, and these are not constant over time among areas. Although all of these factors require consideration, they, too, cannot be taken into account as long as recruiters are allocated strictly, or primarily, on the basis of QMA.

6. ACKNOWLEDGMENTS

We wish to thank Henry Solomon and Marvin Denicoff for their valuable comments, particularly those pertaining to the discussion of the optimal distribution of recruiters. Thanks are also due to the U.S. Marine Corps for help in providing data used in this study. Responsibility for errors in analyzing these data rests with the authors.

BIBLIOGRAPHY

[1] Altman, S., "Earnings, Unemployment and the Supply of Enlisted Volunteers," J. Human Resources *IV*, pp. 38–59 (1969).

[2] Altman, S. and A. C. Fechter, "The Supply of Military Personnel in the Absence of a Draft," Am. Economic Review, pp. 19–31 (May 1967).

[3] Fisher, A. C., "The Cost of the Draft and the Cost of Ending the Draft," Am. Economic Review pp. 239–254 (June 1969).

* Other objectives, such as a geographically representative force, may warrant higher priority. Even for this objective, however, efficiency is of importance when the cost of obtaining necessary force strength or quality of enlistees is judged to be excessively high.

[4] Hansen, W. Lee and B. Weisbrod, "Economics of the Military Draft," Quart. J. Economics, pp. 395–421 (Aug. 1967).

[5] Wool, Harold, *The Military Specialist, Skilled Manpower for the Armed Forces* (The Johns Hopkins Press, 1968).

A BASICALLY POISSON QUEUE WITH NONPOISSON OUTPUT

Nasser Hadidi

Pahlavi University
Shiraz, Iran

ABSTRACT

The output of the queueing system $M/M/1$ is well known to be Poisson. This has also been shown to be true for other more general models inclusive of $M/M_n/1$; the system in which arrivals and epochs of service completion are elements of a birth and death process with parameters λ and $n\mu$, respectively, when the system contains $n \geq 1$ customers. We shall here show that this result is not true in $M_n/M/1$; a system where arrival parameter is state dependent quantity $\lambda/n+1$. Expressions will be given for the steady state joint density of two consecutive output intervals as well as the coefficient of correlation between them.

1. INTRODUCTION

It has been shown by Burke [1] (among others) that the output interval defined as the time interval separating two consecutive departures—after service completion—from the system in the queueing model $M/M/1$ in the steady state follows the inter-arrival interval distribution. Mirasol [7] has shown that this result is also true for the more general model $M/G/\infty$, though his argument needs to be augmented in accordance with Shanbhag [8] and subsequently [9]. This has further been shown, by the author [2], to be true for the model $M/M_n/1$ in which arrivals and epochs of service completion, are elements of a birth and death process with parameters λ and $n\mu$, respectively, when the system contains $n \geq 1$ customers. Other results on such a model are found in Hadidi and Conolly [6] and Hadidi [3] and [4].

Such a model has originally been considered in view of its evident reduction in congestion as explored in the above references. An obvious counterpart to it is the model $M_n/M/1$ where, here, the birth and death process representing epochs of arrival and service completion have parameters $\lambda/n+1$ and μ, respectively, when the system contains $n \geq 0$ customers. Results on this model are found in Hadidi [3] and [5]. We shall show here that, contrary to $M/M_n/1$, the steady state output intervals are not negative exponentially distributed, and consecutive output intervals are stochastically dependent with a coefficient of correlation as given below.

2. PRELIMINARY RESULTS

As in Hadidi [2] we shall need the following:

i) $$Q_r(n) = \Pr[Q(n) = r].$$

where $Q(t)$ represents the state of the system at time t and $Q(n) = Q(t_n + 0)$ with t_n being the epoch of the service completion of the nth customer.

ii) $\gamma_k^{(j)}(s)$; the probability element of a single service interval during which j arrivals take place and starts with $k \geq 1$ customers.

Evidently,

$$\gamma_k^{(0)}(s) = \mu e^{-(\mu+\frac{\lambda}{k+1})s}$$

$$\gamma_k^{(1)}(s) = \int_0^s \frac{\lambda}{k+1} e^{-(\mu+\frac{\lambda}{k+1})t} \mu e^{-(\mu+\frac{\lambda}{k+2})(s-t)}\, dt$$

etc.

Laplace transformation and a subsequent inversion, by standard techniques would produce the general form

$$\gamma_k^{(j)}(s) = \frac{(k+j+1)\mu}{j!} \sum_{i=0}^{j} (-)^{j-i} \binom{j}{i} (k+1+i)^{j-1} e^{-(\mu+\frac{\lambda}{k+1+i})s}. \tag{2.1}$$

Further noting that the Markov chain $Q(n)$ has transition function

$$p_{ij} = \int_0^\infty \gamma_i^{(j+1-i)}(s)\, ds, \qquad j \geq i-1, \qquad i \geq 1$$

$$p_{oj} = \int_0^\infty \int_0^u \lambda e^{-\lambda s} \gamma_1^{(j)}(u-s)\, ds\, du \qquad j \geq 0 \tag{2.2}$$

and

$$p_{ij} = 0 \qquad \text{for } j < i-1,$$

its stationary distribution is given by

$$q_i = \frac{\rho^i}{(i+1)!} q_0;$$

with

$$q_0 = \rho(e^\rho - 1)^{-1}, \qquad \rho = \lambda/\mu. \tag{2.3}$$

We are now in a position to evaluate $h(t, u)$, the steady state joint density of two consecutive output intervals.

3. THE JOINT DENSITY $h(t, u)$

For t and u separating the epoch of completion of service of the $(n-1)$th, nth and $(n+1)$th customers, we distinguish the following five mutually exclusive and exhaustive cases: There are at least two customers left behind after the completion of the $(n-1)$th customer—there is exactly one left and at least one arrival occurs during t—there is exactly one left and no arrival occurs during t—there are no customers left and the nth customer has to wait and finally there are no customers left and the nth customer finds the system empty on his arrival.

Adding the contribution of these cases in the steady state, we obtain:

$$
\begin{aligned}
(3.1)\qquad h(t,u) = (e^{\rho}-1)^{-1}\Bigg[& \sum_{i\geq 2}\frac{\rho^{i+1}}{(i+1)!}\mu e^{-\mu t}\,\mu e^{-\mu u} + \frac{\rho^2}{2!}\left\{\mu e^{-\mu t} - \mu e^{-\left(\mu+\frac{\lambda}{2}\right)t}\right\}\mu e^{-\mu u} \\
& + \frac{\rho^2}{2!}\mu e^{-\left(\mu+\frac{\lambda}{2}\right)t}\int_0^u \lambda e^{-\lambda x}\,\mu e^{-\mu(u-x)}\,dx \\
& + \frac{\rho}{1!}\int_0^t \lambda e^{-\lambda x}\left\{\mu e^{-\mu(t-x)} - \mu e^{-\left(\mu+\frac{\lambda}{2}\right)(t-x)}\right\}\mu e^{-\mu u}\,dx \\
& + \frac{\rho}{1!}\int_0^t \lambda e^{-\lambda x}\,\mu e^{-\left(\mu+\frac{\lambda}{2}\right)(t-x)}\,dx\int_0^u \lambda e^{-\lambda y}\,\mu e^{-\mu(u-y)}\,dy\Bigg],
\end{aligned}
$$

which can be simplified to

$$
\begin{aligned}
(3.2)\qquad h(t,u) = \mu^2 e^{-\mu(t+u)}\Bigg[1 + \frac{\rho}{(e^{\rho}-1)(\rho-1)}\Bigg[1 + \frac{\rho}{\rho-2}\Bigg[& \left(\frac{\rho}{2}+1\right)e^{-\frac{\lambda}{2}t}\left\{1-\rho e^{-(\lambda-\mu)u}\right\} \\
& - \rho e^{-(\lambda-\mu)t}\left\{1-2e^{-(\lambda-\mu)u}\right\}\Bigg]\Bigg]\Bigg]
\end{aligned}
$$

The marginal density of t is

$$
(3.3)\qquad g(t) = \mu e^{-\mu t} + \frac{\rho}{(\rho-1)(e^{\rho}-1)}\left(\mu e^{-\mu t} - \lambda e^{-\lambda t}\right)
$$

with mean

$$
(3.4)\qquad E(t) = \frac{e^{\rho}}{\mu(e^{\rho}-1)},
$$

as expected (see Hadidi [5] where the mean inter-arrival interval is shown to be (3.4) in the steady state) and

$$
(3.5)\qquad \mathrm{Var}(t) = \frac{1}{\mu^2}\left[1 + \frac{1}{e^{\rho}-1}\left(\frac{2}{\rho} - \frac{1}{e^{\rho}-1}\right)\right].
$$

The coefficient of correlation between t and u is

$$r = \frac{e^{\rho}(2-\rho^2) - 2(1+\rho)}{[\rho e^{2\rho} + 2e^{\rho}(1-\rho) - 2](\rho+2)}, \tag{3.6}$$

which amounts to

ρ	r
0.2	−0.012
0.5	−0.045
0.9	−0.076
0.95	−0.078

REFERENCES

[1] Burke, P. J., "Output of Queueing Systems," Oprs. Res *4* 699–707 (1956).

[2] Hadidi, N., "On the Output Process of a State Dependent Queue," Skand. AktuarTidskr., pp. 182–186 (1972).

[3] Haddi, N., "On the Busy Period of a State-Dependent Queue," to appear J. Appl. Prob. (1974).

[4] Hadidi, N., "On the Service Time Distribution and Waiting Time Process of Potentially Infinite Capacity Queueing System," J. Appl. Prob. *6*, 594–603 (1969).

[5] Hadidi, N., "A Queueing Model with Variable Arrival Rates," to appear, Periodica Mathematica Hungarica (1973).

[6] Hadidi, N. and B. W. Conolly, "On the Improvement of the Operational Characteristics of Single Server Queues by Use of Queue Length Dependent Service Mechanism," J. Roy. Statist. Soc. (Series C) *18*, 229–240 (1969).

[7] Mirasol, N. M. "The Output of $M/G/\infty$ is Poisson," Operations Research *11*, 282–284 (1963).

[8] Shanbhag, D. N. "Letter to the Editor", J. Appl. Prob. *9*, 470 (1972).

[9] Shanbhag, D. N. and D. G. Tambouratzis, "Erlang's Formula and some Results on the Departure Process for a Loss system," J. Appl. Prob. *10*, 233–240 (1973).

OPTIMIZATION OF STRATEGIC DEFENSES TO PROVIDE SPECIFIED POST-ATTACK PRODUCTION CAPACITIES

Jerome Bracken and James T. McGill*

Institute for Defense Analyses
Arlington, Virginia

ABSTRACT

This paper presents a model for choosing a minimum-cost mix of strategic defenses to assure that specified production capacities for several economic sectors survive after a nuclear attack. The defender selects a mix of strategic defenses for each of several geographic regions. The attacker chooses an allocation of attacking weapons to geographic regions, within specified weapon inventories. The attack is optimized against any economic sector. This formulation allows the defense planner the capability to assess the results of the optimal defense structure for a "worst case" attack. The model is a mathematical program with nonlinear programming problems in the constraints; an example of its application is given and is solved using recently developed optimization techniques.

MODEL

One of the planning considerations for defense against a nuclear attack is the assurance of sufficient surviving economic capacity to support the surviving population. This paper uses an aggregate model of the nationwide production base. The effect of strategic defenses, consisting of both active defenses (for example, antiballistic missiles) and passive defenses (for example, population shelters), against offensive weapons, is modeled. The defender chooses a minimum-cost mix of active and passive components which ensures that specified post-attack production capacities will survive after an optimized attack.

The country is partitioned into $j=1, \ldots, n$ geographic regions, allowing consideration of varying population and production-base densities. The general model allows $i=1, \ldots, m$ different economic sectors in each geographic region. Each economic sector is characterized by a Cobb-Douglas production function. Also, $k=1, \ldots, p$ denotes the different types of defensive resources and $l=1, \ldots, q$ the different types of offensive weapons.

Define

x_{jk} = number of defensive resources of type k assigned to region j,

v_{jl}^{i} = number of offensive resources of type l targeted on region j in an attack on economic sector i,

V_l = number of offensive weapons of type l.

The post-attack production function (in terms of value added) in economic sector i in region j is

$$H_{ij}\left[1-e^{-\sum_{k=1}^{p} a_{ijk}^{H} x_{jk}} \left(1-e^{-\sum_{l=1}^{q} b_{ijl}^{H} v_{jl}^{i}}\right)\right] \cdot K_{ij}^{\alpha_{ij}}\left[1-e^{-\sum_{k=1}^{p} a_{ijk}^{K} x_{jk}} \left(1-e^{-\sum_{l=1}^{q} b_{ijl}^{K} v_{jl}^{i}}\right)\right]^{\alpha_{ij}} \cdot L_{ij}^{\beta_{ij}}\left[1-e^{-\sum_{k=1}^{p} a_{ijk}^{L} x_{jk}} \left(1-e^{-\sum_{l=1}^{q} b_{ijl}^{L} v_{jl}^{i}}\right)\right]^{\beta_{ij}}.$$

*Now with Illinois Board of Higher Education, Springfield, Illinois.

If $v_{jl}^i = 0, l = 1, \ldots, q$, then the production function has the standard Cobb-Douglas form

$$H_{ij} K_{ij}^{\alpha_{ij}} L_{ij}^{\beta_{ij}},$$

where H_{ij} represents the technological efficiency, K_{ij} is the capital base, and L_{ij} is the labor base. The exponents α_{ij} and β_{ij} are the elasticities of value added with respect to capital and labor, respectively.

The expressions of the form

$$1 - e^{-\sum_{k=1}^{p} a_{ijk} x_{jk}} \qquad \left(1 - e^{-\sum_{l=1}^{q} b_{ijl} v_{jl}^i}\right)$$

modify the efficiency, the capital base, and the labor base as a function of the offense and defense allocations. The expression is convex in v_{jl}^i, concave in x_{jk}, and has the asymptotic properties expected for the physical processes being modeled. The parameters a_{ijk} and b_{ijl} might be estimated from detailed analyses. One means of deriving these estimates is given in the next section.

The cost of defensive resources is taken to be $\sum_{k=1}^{p} \sum_{j=1}^{n} c_{jk} x_{jk}$, where c_{jk} is the unit cost of defensive resource k in region j.

The model given below requires that the surviving nationwide post-attack production capacity for economic sector i be specified to be at least $r^i (i = 1, \ldots, m)$. The aggregate production for the whole country is the sum of the production in all geographic regions. Alternatively, production survival requirements could be specified by region rather than aggregated.

The overall model is to choose $x_{jk} (j = 1, \ldots, n;\ k = 1, \ldots, p)$ and $v_{jl}^i (i = 1, \ldots, m;\ j = 1, \ldots, n;\ l = 1, \ldots, q)$ to

$$\text{minimize} \sum_{k=1}^{p} \sum_{j=1}^{n} c_{jk} x_{jk}$$

subject to

$$\left[\begin{array}{l} \underset{v_{jl}^i}{\text{minimum}} \sum_{j=1}^{n} H_{ij} \left[1 - e^{-\sum_{k=1}^{p} a_{ijk}^H x_{jk}}\right] \left[1 - e^{-\sum_{l=1}^{q} b_{ijl}^H v_{jl}^i}\right]^{\alpha_{ij}} \\ \quad \cdot K_{ij}^{\alpha_{ij}} \left[1 - e^{-\sum_{k=1}^{p} a_{ijk}^K x_{jk}}\right] \left[1 - e^{-\sum_{l=1}^{q} b_{ijl}^K v_{jl}^i}\right]^{\beta_{ij}} \\ \quad \cdot L_{ij}^{\beta_{ij}} \left[1 - e^{-\sum_{k=1}^{p} a_{ijk}^L x_{jk}}\right] \left[1 - e^{-\sum_{l=1}^{q} b_{ijl}^L v_{jl}^i}\right] \end{array} \geqslant r^i, \qquad i = 1, \ldots, m.\right]$$

$$\sum_{j=1}^{n} v_{jl}^i \leqslant V_l, \qquad l = 1, \ldots, q$$

PARAMETERS FOR APPLICATION OF MODEL

The following dimensions of the problem are used in the application.

(1) aggregate industries
  $i = 1$ agriculture and service
  $i = 2$ manufacturing
(2) geographic regions
  $j = 1$ East
  $j = 2$ West
  $j = 3$ South
(3) defenses
  $k = 1$ active
  $k = 2$ passive
(4) offensive weapons
  $l = 1$ 1-megaton weapon

Values for each of the parameters are developed below. The assumptions used in the derivation of these values may be changed or relaxed by a user to reflect a perhaps more realistic description of the physical and economic processes embodied in the model. The approach taken herein is to make relatively simple parametric assumptions to illustrate the application.

Economic Parameters

An economic data base developed for the Defense Civil Preparedness Agency and described in Reference [5] is used in estimating the parameters of the aggregate production functions. That data base has 82 economic sectors. The first 77 are used here, aggregated to an "agriculture and service" industry and to a "manufacturing" industry for each of three regions—East (roughly, those states east of the Mississippi River and north of the Mason-Dixon line), West (west of the Mississippi River), and South (east of the Mississippi River and south of the Mason-Dixon line). The total 1970 population is about 202 million with 46, 32, and 22 percent being in the East, West, and South, respectively.

The capital stock, K, is aggregated from individual state and economic sector estimates to the aggregate industries and regions used here. The total labor force is taken to be 44 percent of each region's population. The total is distributed across the two industries by equating the marginal products of labor in each. The parameters α and β are weighted averages of the parameters estimated for each of the original 77 sectors [4]. The weighting factors are the values of capital stocks. Finally, the technological efficiency parameter, H, is determined by requiring that estimated output using K, L, α, and β equals actual output. Table 1 summarizes total value added by industry and region and Table 2 gives the production function parameter estimates. The total value added for the United States is \$866 billion, with \$555 billion in agriculture and services and \$311 billion in manufacturing.

Offense Parameters

Assuming that production technological efficiency in the short run is not affected by the attack,

$$b^H_{ij1} = 0, \qquad i = 1, 2; \qquad j = 1, 2, 3.$$

Further assuming that capital and labor are destroyed identically within industries and geographic regions,

$$b_{ij1}^{K} = b_{ij1}^{L}, \qquad i = 1, 2; \qquad j = 1, 2, 3.$$

To estimate these parameters, attack optimizations are made against undefended targets using a procedure described in Reference [5]. Six attack optimizations are made, for each of the two industries in each of the three regions. The aggregate industries and geographic regions include the individual industries and states as described previously. The measure used in the optimization is total value added. Thus, each of the offense effectiveness parameters is estimated under the assumption that the offense targets its weapons to maximize damage to the production base for the industry being attacked.

TABLE 1. *Value added*

Industry	Regions			Total
	1 (East)	2 (West)	3 (South)	
1.....	0.241×10^{12}	0.203×10^{12}	0.111×10^{12}	0.555×10^{12}
2.....	0.168×10^{12}	0.091×10^{12}	0.052×10^{12}	0.311×10^{12}

TABLE 2. *Production function parameters*

Industry		Regions		
		1	2	3
1.....	H	7.87	4.98	4.42
	K	0.663×10^{12}	0.598×10^{12}	0.329×10^{12}
	α	0.705	0.746	0.752
	L	20.1×10^{6}	15.4×10^{6}	10.5×10^{6}
	β	0.295	0.254	0.248
2.....	H	0.0075	0.0138	0.0194
	K	0.809×10^{12}	0.374×10^{12}	0.228×10^{12}
	α	0.844	0.823	0.815
	L	21.2×10^{6}	12.7×10^{6}	9.1×10^{6}
	β	0.446	0.466	0.454

The weapons used are 1 megaton, with assumed overall probability 0.75 of successfully impacting in the target area.

Table 3 displays the results of an optimized attack on each industry in each region for 1,000 and for 2,000 weapons. The offense weapon effectiveness parameter, b, is computed by letting

$$1 - e^{-bv}$$

equal the fraction of total value added that is destroyed by v weapons. The values for each industry and region and for 1,000 and 2,000 weapons are given in Table 4.

TABLE 3. *Fraction of total value added destroyed*

Aggregate industry	Region	Weapon inventory	
		1,000	2,000
1	East	0.605	0.790
1	West	0.572	0.722
1	South	0.697	0.821
2	East	0.563	0.720
2	West	0.642	0.795
2	South	0.692	0.813

TABLE 4. *Values of* "b" *for weapon inventories of 1,000 and 2,000*

Aggregate industry	Region	Weapon inventory	
		1,000	2,000
1	East	0.00093	0.00078
1	West	0.00085	0.00064
1	South	0.00120	0.00086
2	East	0.00083	0.00064
2	West	0.00103	0.00079
2	South	0.00118	0.00084

Defense Parameters

Active defenses consist of interceptors and associated radars. Passive defenses consist of blast shelters, with one unit of shelter defense taken to represent 10,000 people sheltered.

Active Defense.—Since short run production technological efficiency is not affected by the attack, a^H_{ij1} can be taken to be 0 for $i=1, 2$; $j=1, 2, 3$.

Assuming that capital stock and population are protected equally by the active defense,

$$a^K_{ij1} = a^L_{ij1}, \qquad i=1, 2; \qquad j=1, 2, 3.$$

By using the offense effectiveness parameters, b, displayed in Table 4, the number of offensive weapons required to destroy half of the value for each industry and geographic region is computed and shown in Table 5. The active defense effectiveness parameters, a, are computed on the basis that 1.5 interceptors would be required to protect half the value were the offense to target all of his weapons on the particular industry and geographic region. Thus, for instance, 255 weapons would have to be intercepted in a 1,000 weapon attack on industry 1 in the East, requiring 383 interceptors. The required interceptors are given in Table 6.

The active defense effectiveness parameters are computed from

$$1-e^{-ax}=0.5,$$

where x is the value in Table 6, resulting in the values given in Table 7.

Passive Defense.—Since short run production technological efficiency is not affected by the attack, a^H_{ij2} can be taken to be 0 for $i=1, 2$; $j=1, 2, 3$.

It is assumed that shelters do not protect capital stock, resulting in

$$a^{K}_{ij2} = 0, \quad . \quad i = 1, 2; \qquad j = 1, 2, 3.$$

TABLE 5. *Weapons required to destroy half of value*

Aggregate industry	Region	Weapon inventory	
		1,000	2,000
1	East	745	889
1	West	815	1083
1	South	578	806
2	East	835	1083
2	West	673	877
2	South	587	825

TABLE 6. *Interceptors required to protect half of value*

Aggregate industry	Region	Weapon inventory	
		1,000	2,000
1	East	383	1667
1	West	278	1376
1	South	633	1791
2	East	248	1376
2	West	491	1685
2	South	620	1763

TABLE 7. *Values of* "a" *for Weapon Inventories of 1,000 and 2,000*

Aggregate industry	Region	Weapon inventory	
		1,000	2,000
1	East	0.00181	0.00042
1	West	0.00249	0.00050
1	South	0.00110	0.00039
2	East	0.00279	0.00050
2	West	0.00141	0.00041
2	South	0.00112	0.00039

The urban population is taken to be 150 million people distributed across regions in proportion to total population. Then, urban population in East, West and South is 69.5 million, 47.5 million, and 33 million, respectively. Shelter units in numbers 6950, 4750, and 3300 by region will protect the whole urban population. If half this many shelters will protect half the urban population, shelter effectiveness values can be calculated such that

$$1 - e^{-ax} = 0.5.$$

Taking x to be 3475, 2375, and 1650 for each of the three regions results in

$$a^L_{i12} = 0.00020$$
$$a^L_{i22} = 0.00029 \qquad 1 = 1, 2.$$
$$a^L_{i32} = 0.00042$$

Costs of Active and Passive Defense

Cost per interceptor is taken to be \$3 million, and cost per radar \$400 million. Assuming that one radar is needed for every 80 interceptors, total cost per interceptor is \$8 million. Thus

$$c_{j1} = 8, \qquad j = 1, 2, 3.$$

Cost per shelter space is taken to be \$200. The total cost per 10,000 people sheltered is \$2 million. Thus

$$c_{j2} = 2, \qquad j = 1, 2, 3.$$

SUMMARY OF MODEL INCLUDING PARAMETERS

The complete model for a 1,000-weapon attack is given below. The offense allocates its weapons to the East, West, and South to minimize productive capacity in either industry 1 or industry 2. The effect of a weapon is to destroy capital stock and the effective labor force, thus reducing the ability of the defense to generate output in the two industries. The defense chooses numbers of interceptors and shelters in each geographic region to guarantee that total productive capacity is at least half of the pre-attack level, no matter on which industry the offense concentrates his attack. The surviving productive capacity would in reality apply only if the necessary distribution systems remained sufficiently intact. Future extensions of the model might include explicit consideration of such systems.

The model is to choose $x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, x_{32}$; $v^1_{11}, v^1_{21}, v^1_{31}, v^2_{11}, v^2_{21}, v^2_{31}$ to

$$\text{minimize } 8x_{11} + 2x_{12} + 8x_{21} + 2x_{22} + 8x_{31} + 2x_{32}$$

subject to

$$\begin{aligned}
\underset{v^1_{j1}}{\text{minimum}} \quad & 7.87\left[1 - e^{-0x_{11} - 0x_{12}} \quad (1 - e^{-0v^1_{11}})\right] \\
& \cdot (0.663 \times 10^{12})^{0.705} \left[1 - e^{-0.00181x_{11} - 0x_{12}} \quad (1 - e^{-0.00093v^1_{11}})\right]^{0.705} \\
& \cdot (20.1 \times 10^{6})^{0.295} \left[1 - e^{-0.00181x_{11} - 0.00020x_{12}} \quad (1 - e^{-0.00093v^1_{11}})\right]^{0.295} \\
+ \quad & 4.98\left[1 - e^{-0x_{21} - 0x_{22}} \quad (1 - e^{-0v^1_{21}})\right] \\
& \cdot (0.598 \times 10^{12})^{0.746} \left[1 - e^{-0.00249x_{21} - 0x_{22}} \quad (1 - e^{-0.00085v^1_{21}})\right]^{0.746} \\
& \cdot (15.4 \times 10^{6})^{0.254} \left[1 - e^{-0.00249x_{21} - 0.00029x_{22}} \quad (1 - e^{-0.00085v^1_{21}})\right]^{0.254} \\
+ \quad & 4.42\left[1 - e^{-0x_{31} - 0x_{32}} \quad (1 - e^{-0v^1_{31}})\right] \\
& \cdot (0.329 \times 10^{12})^{0.752} \left[1 - e^{-0.00110x_{31} - 0x_{32}} \quad (1 - e^{-0.00120v^1_{31}})\right]^{0.752} \\
& \cdot (10.5 \times 10^{6})^{0.248} \left[1 - e^{-0.00110x_{31} - 0.00042x_{32}} \quad (1 - e^{-0.00120v^1_{31}})\right]^{0.248}
\end{aligned} \geqslant 0.2775 \times 10^{12}$$

$$v^1_{11} + v^1_{21} + v^1_{31} \leqslant 1{,}000$$

$$
\begin{aligned}
\underset{v_{j1}^2}{\text{minimum}} \quad & 0.00746[1 - e^{-0x_{11} - 0x_{12}} \quad (1 - e^{-0v_{11}^2})] \\
& \cdot (0.809 \times 10^{12})^{0.844} [1 - e^{-0.00279x_{11} - 0x_{12}} \quad (1 - e^{-0.00083v_{11}^2})]^{0.844} \\
& \cdot (21.2 \times 10^{6})^{0.446} \quad [1 - e^{-0.00279x_{11} - 0.00020x_{12}} (1 - e^{-0.00083v_{11}^2})]^{0.446} \\
+ \quad & 0.0138[1 - e^{-0x_{21} - 0x_{22}} \quad (1 - e^{-0v_{21}^2})] \\
& \cdot (0.374 \times 10^{12})^{0.823} [1 - e^{-0.00141x_{21} - 0x_{22}} \quad (1 - e^{-0.00103v_{21}^2})]^{0.823} \\
& \cdot (12.7 \times 10^{6})^{0.466} \quad [1 - e^{-0.00141x_{21} - 0.00029x_{22}} (1 - e^{-0.00103v_{21}^2})]^{0.466} \\
+ \quad & 0.0194[1 - e^{-0x_{31} - 0x_{32}} \quad (1 - e^{-0v_{31}^2})] \\
& \cdot (0.228 \times 10^{12})^{0.815} [1 - e^{-0.00112x_{31} - 0x_{32}} \quad (1 - e^{-0.00118v_{31}^2})]^{0.815} \\
& \cdot (9.09 \times 10^{6})^{0.454} \quad [1 - e^{-0.00112x_{31} - 0.00042x_{32}} (1 - e^{-0.00118v_{31}^2})]^{0.454}
\end{aligned}
\geqslant 0.1555 \times 10^{12}
$$

$$v_{11}^2 + v_{21}^2 + v_{31}^2 \leqslant 1{,}000$$

SOLUTION PROCEDURE

Reference [1] shows that mathematical programs with nonlinear programs in the constraints can be convex programs if certain conditions hold. The model presented in this paper satisfies the conditions. Reference [2] documents a computer program for solving mathematical programs with nonlinear programs in the constraints. It is called INSUMT, and is used with the most recent SUMT program documented in Reference [6].

User-supplied subroutines for use with SUMT and INSUMT were programmed for the application of the model. These subroutines are available from the authors.

It should be noted that the value of a nonlinear program in the constraints is not necessarily a continuously differentiable function of the variables of the overall program, and hence the solution procedure cannot be guaranteed to converge to a global solution. However, use of alternative starting points can provide confidence that a global solution is attained.

RESULTS

Tables 8 and 9 display the minimum-cost mixes of active and passive defenses for post-attack production capacity requirements of 75 and 90 percent of pre-attack capacity. Also shown is the optimal targeting of the offense against the chosen defenses for each of the two industries. No defenses are required in the 50-percent case. A significant number of shelters are necessary for the 75-percent requirement. The defense is predominantly active for the 90-percent case.

TABLE 8. *Results for 1,000 weapons and 75-percent survival requirements*

Region	Defense			Offense	
	Interceptors	Shelters	Cost	Weapons targeted on industry	
				1	2
East	224	620	3.0	626	517
West	238	326	2.6	170	410
South	238	9	1.9	202	72
Total	700	955	7.5	998	999

TABLE 9. *Results for 1,000 weapons and 90-percent survival requirements*

Region	Defense			Offense	
	Interceptors	Shelters	Cost	Weapons targeted on industry	
				1	2
East	1896	5	15.2	3	1
West	1317	82	10.7	2	322
South	738	45	6.0	994	676
Total	3951	132	31.9	999	999

Tables 10 and 11 give minimum-cost defense mixes for 2,000 offense weapons at 50- and 75-percent surviving capacity requirements. The defenses are primarily active for both cases, although the amount of shelters is substantially greater in geographic region 3.

TABLE 10. *Results for 2,000 weapons and 50-percent survival requirements*

Region	Defense			Offense	
	Interceptors	Shelters	Cost	Weapons targeted on industry	
				1	2
East	100	1	0.8	1081	1197
West	100	0	0.8	723	690
South	99	20	0.8	195	112
Total	299	21	2.4	1999	1999

TABLE 11. *Results for 2,000 weapons and 75-percent survival requirements*

Region	Defense			Offense	
	Interceptors	Shelters	Cost	Weapons targeted on industry	
				1	2
East	2051	4	16.4	889	876
West	1897	31	15.2	335	615
South	355	1057	5.0	773	509
Total	4303	1092	36.6	1997	2000

ACKNOWLEDGMENTS

This work draws upon results of an economic model developed by Hugh M. Pitcher and an attack optimization procedure developed by Leo A. Schmidt. Carolyn D. Davis provided helpful assistance. Garth P. McCormick consulted on solution of the problem.

REFERENCES

[1] Bracken, J. and J. T. McGill, "Mathematical Programs With Optimization Problems in the Constraints," Operations Research *21*, 37–44 (1973).

[2] Bracken, J. and J. T. McGill, Computer Program for Solving Mathematical Programs With Nonlinear Programs in the Constraints, Institute for Defense Analyses, P–801 (Mar. 1972) (available as document number AD 742 098, National Technical Information Service, Springfield, Va. 22151). (See also Bracken, J. and McGill, J.T., A Method for Solving Mathematical Programs With Optimization Problems in the Constraints, to appear in Operations Research.)

[3] Fiacco, A. V. and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* (Wiley, N.Y., 1968).

[4] Grimm, B. T., Estimation of CES Production Functions for U.S. Manufacturing by Input-Output Sector, Institute for Defense Analyses, P–525 (July 1969).

[5] McGill, J. T., et al., Methodologies for Evaluating Vulnerability of National Systems, Institute for Defense Analyses, S–394 (June 1972).

[6] Mylander, W. C., R. L. Holmes, and G. P. McCormick, A Guide to SUMT—Version 4: The Computer Program Implementing the Sequential Unconstrained Minimization Technique, Research Analysis Corporation, P–63 (Oct. 1971) (available as document number AD 731 391, National Technical Information Service, Springfield, Va. 22151).

COMPARISON OF STOCHASTIC PROCESSES USED IN SONAR DETECTION MODELS*

Bernard J. McCabe

Daniel H. Wagner, Associates
Paoli, Pennsylvania

ABSTRACT

In any model for a sonar detection process, some assumption must be made about the nature of the acoustic fluctuation process. Two processes that are widely used in this role are the jump process and the Gauss-Markov process. These processes are similar in that they are both stationary Markov processes and have autocovariance functions of the form $\sigma^2\exp(-\gamma|t|)$. For these reasons, it might be believed that one could use either of these processes and get comparable results if all one is interested in is computing cumulative detection probabilities or mean time to gain or lose contact. However, such is not the case in that vastly different results can be obtained in some applications. An application of this sort is presented. We also present necessary and sufficient conditions for a threshold to have the property that it is almost surely crossed by the jump process, or by the Gauss-Markov process. This affords another method of comparison.

1. INTRODUCTION

In the operations analysis of submarine vs submarine passive sonar detection problems one often models this detection process by assuming that contact is held on the target at time t if signal excess measured in decibels, $S(t)$, is nonnegative; for example see [11], p. 20. In reconstructing submarine vs submarine encounters, however, it is often found that detection occurs or contact is held even though the estimated signal excess is negative, and that sometimes detection fails to occur or contact is not held even though signal excess is positive. Thus it is convenient to hypothesize that the actual signal excess is given by an expression of the form

$$S(t) = M(t) + \xi(t)$$

where M is a deterministic function given by the sonar equation and depends on the target's radiated noise, receiver's self-noise, and the propagation loss suffered between target and receiver, and ξ is a mean-zero stochastic process which represents the unknown random fluctuations in the above quantities. It is then assumed that the receiving submarine holds contact on the target submarine at time t if and only if

* This work was performed for Commander Submarine Development Group Two, Groton, Connecticut, and was supported by the Office of Naval Research under Contract No. N00014-70-C-0261, and under Contract No. N00014-73-C-0432.

$$S(t) = M(t) + \xi(t) \geq 0, \tag{1}$$

and the initial detection takes place at the first time t at which (1) holds.

This model for detection, with a variety of choices for the random process ξ, is used in a large number of Naval sponsored simulation programs (APSUB, SIM II, the Submarine Development Group Two simulation model) and operations analysis studies ([12], for example).

2. STOCHASTIC MODELS FOR FLUCTUATIONS

As for the random process ξ there are two processes which have enjoyed almost exclusive use in this context: the jump process ξ_J and the Gauss-Markov process ξ_G. The jump process is defined as follows: let $N(t)$, $t \geq 0$, be a Poisson process with intensity $\lambda \geq 0$ and let $\{Z_0, Z_1, \ldots\}$ be a sequence of independent normally distributed random variables with mean 0 and variance σ^2 and let

$$\xi_J(t) = Z_{N(t)}.$$

In words, ξ_J changes values, to a new independent draw from a normal distribution at instants which are separated by an exponentially distributed amount of time with mean $1/\lambda$. This process was originally used in this context by J. D. Kettelle (see [8]) and was considered to be a representation of the rapid surges and fades in acoustic conditions described by sonar operators and discerned in some bodies of sonar data.

It is also a convenient model mathematically because most quantities of interest such as cumulative detection probability,

$$CDP(t) = P(M(x) + \xi_J(x) \geq 0 \qquad \text{for some } x \leq t),$$

can be computed analytically for rather general mean signal excess histories M without resorting to Monte Carlo simulations (see [2] for the continuous case and [4] for the discrete case).

The Gauss-Markov process ξ_G can be defined as the stationary Gaussian process which is also a (strict sense) Markov process and has autocovariance function

$$\begin{aligned} \psi(t) &= E\xi_G(s)\xi_G(s+t) \\ &= \sigma^2 e^{-\lambda|t|} \end{aligned} \tag{2}$$

for some $\sigma^2 \neq 0$, $\lambda > 0$. (See [5] and [6] for further properties of this process.) B. O. Koopman [9] has argued from an information theoretic viewpoint that the Gauss-Markov process is in a sense the most natural process to use in this context. It is available for use in many of the commonly used ASW simulations, but has the disadvantage that very few analytic computations can be made using it as a model for the random component of signal excess. Thus one commonly has to resort to Monte Carlo methods. Reference [4] contains substantial material concerning the use of the Gauss-Markov process and the jump process in detection models.

Now it happens that the jump process and the Gauss-Markov process share four important properties: they are both stationary and Markov (strict sense) processes, their autocovariance functions are

identical and given by (2), and their one-dimensional distributions are Gaussian. Some important differences between them are: the Gauss-Markov process has continuous (although nondifferentiable) sample paths and has Gaussian finite-dimensional distributions, while the jump process has step functions for its sample paths and its d-dimensional distributions are non-Gaussian $d \geq 2$.

However, because of the similarities of these two processes it might be thought that whichever of them were used one would obtain comparable numerical results if one were computing such quantities as cumulative detection probability or mean time to lose or gain contact. The purpose of this paper is to show that such is not the case; the choice of which process to use is crucial in that vastly different results may be obtained by using one rather than the other. This is not to say that either of these processes is the correct one to use, but rather to dramatize the importance of knowing which process is more faithful to reality: even a reasonably close match may not be good enough for drawing conclusions in some ASW applications.

3. Mean Time to Lose Contact

To illustrate the contention that ξ_G and ξ_J as described above yield widely disparate answers to some questions consider the following application. Suppose a tracking submarine maneuvers so as to maintain a constant range from a target submarine and that at this range a mean signal excess of M decibels persists, $M > 0$. Thus contact will be maintained until the first time t that $M + \xi(t) < 0$, and we assume that $\xi(0) = 0$.

As a measure of how well the tracking submarine performs in this role we consider the quantity: mean time to lose contact. Thus let $\mu_J(M, \lambda, \sigma)$ be the mean time to lose contact given that $\xi = \xi_J$ with parameters λ, σ and that the initial signal excess is M decibels, and let $\mu_G(M, \lambda, \sigma)$ be the same quantity but with $\xi = \xi_G$.

In the case $\xi = \xi_J$ the probability distribution of time to lose contact can be computed. In fact,

P(loss of contact occurs after time t)

$$= \sum_{n=0}^{\infty} P(\text{loss of contact after } t \text{ and } n \text{ jumps occur before } t)$$

$$= \sum_{n=0}^{\infty} P(\text{loss of contact after } t | n \text{ jumps before } t)\, P(n \text{ jumps before } t)$$

$$= \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} p^n$$

$$= e^{-\lambda t(1-p)},$$

where

$$p = \frac{1}{\sqrt{2\pi}} \int_{-M/\sigma}^{\infty} e^{-x^2/2} dx.$$

Thus

$$\mu_J(M, \lambda, \sigma) = \int_0^{\infty} e^{-\lambda t(1-p)} dt \tag{3}$$

$$= \frac{1}{\lambda(1-p)}$$

$$= \left(\frac{\lambda}{\sqrt{2\pi}} \int_{M/\sigma}^{\infty} e^{-t^2/2} dt \right)^{-1}$$

and using a familiar approximation ([6], p. 166),

$$\mu_J(M, \lambda, \sigma) \sim \frac{\sqrt{2\pi}}{\lambda} e^{M^2/2\sigma^2} \frac{M}{\sigma}. \tag{4}$$

(The tilde notation, here and below, means that the ratio of the two sides tends to 1 as $M \to \infty$.)

However, as will be shown in the next section,

$$\mu_G(M, \lambda, \sigma) \sim \frac{\sqrt{2\pi}}{\lambda} e^{M^2/2\sigma^2} \frac{\sigma}{M}. \tag{5}$$

Comparing (4) and (5) one sees that

$$\mu_J(M, \lambda, \sigma) \sim \mu_G(M, \lambda, \sigma) \frac{M^2}{\sigma^2} \tag{6}$$

Thus, for example, if $\sigma = 9$ decibels and $M = 18$ decibels these two estimates of mean time differ by a multiplicative factor of four. Figure 1 shows the two curves μ_G and μ_J using the exact formulae (3) and (7) below.

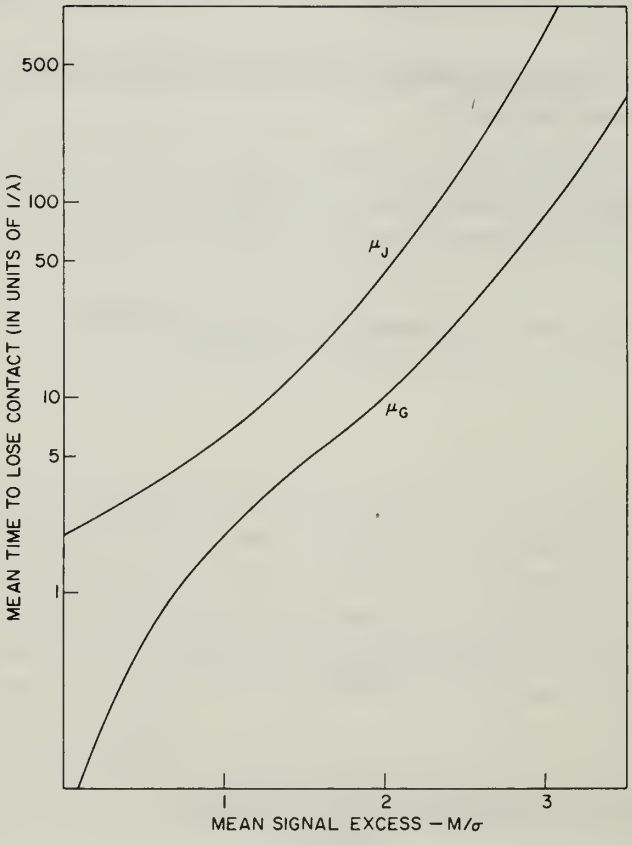


FIGURE 1. Mean time to lose contact for jump process and Gauss-Markov process.

Obviously it is difficult to assess the capabilities of a submarine in this role without knowing which of these processes is a correct model for acoustic fluctuations (if indeed either of them is). Unfortunately that is just the position in which we find ourselves.

4. DERIVATION OF THE ESTIMATE (5)

The derivation of (5) has been made possible by the recent appearance of a result due to Belkin and Snyder [3]. They prove that, in our notation,

(7) $$\mu_G(M, \lambda, \sigma) = \frac{\sqrt{\pi}}{\lambda} \sum_{n=1}^{\infty} \frac{(M/\sigma)^n}{2^{n/2} n \Gamma\left(\frac{n+1}{2}\right)}.$$

What we need is an asymptotic formula which displays the dependence of μ_G on M in a more transparent way, so that a comparison to μ_J can more readily be made.

To derive the estimate (5) we use the following collection of definitions and relations:

(A) $$\operatorname{erf}(x) = 1 - \operatorname{erfc}(x)$$

$$= \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

(B) $$\omega(z) = e^{-z^2} \operatorname{erfc}(-iz)$$

(C) $$= \sum_{n=0}^{\infty} \frac{(iz)^n}{\Gamma\left(\frac{n}{2}+1\right)} \qquad \text{(Reference [1], p. 297)}$$

(D) $$e^{x^2} \int_x^{\infty} e^{-t^2} dt \leq (x + \sqrt{x^2 + 4/\pi})^{-1} \qquad \text{(Reference [1], p. 298)}$$

(E) $$\int_0^x e^{t^2/2}\, dt \sim x^{-1} e^{x^2/2} \qquad \text{(Reference [10])}.$$

Starting with the identity (7) then, and setting $L = M/\sigma$, we have

$$\mu_G(M, \lambda, \sigma) = \frac{\sqrt{\pi}}{\lambda} \sum_{n=1}^{\infty} \frac{(L/\sqrt{2})^n}{n \Gamma\left(\frac{n+1}{2}\right)}$$

$$= \lambda^{-1} \sqrt{\frac{\pi}{2}} \int_0^L \sum_{n=0}^{\infty} \frac{(x/\sqrt{2})^n}{\Gamma\left(\frac{n}{2}+1\right)}\, dx$$

$$= \lambda^{-1} \sqrt{\frac{\pi}{2}} \int_0^L \omega(-ix/\sqrt{2})dx$$

$$= \lambda^{-1} \sqrt{\frac{\pi}{2}} \int_0^L e^{x^2/2} \operatorname{erfc}\ (-x/\sqrt{2})dx$$

$$= \lambda^{-1} \sqrt{\frac{\pi}{2}} \int_0^L e^{x^2/2}(2 - \operatorname{erfc}\ (x/\sqrt{2}))dx$$

$$= \lambda^{-1}\ \sqrt{2\pi} \int_0^L e^{x^2/2}dx - O\ (\log L) \qquad \text{(using (D) above)}$$

$$\sim \frac{\sqrt{2\pi}}{\lambda} e^{L^2/2} \frac{1}{L}$$

$$= \frac{\sqrt{2\pi}}{\lambda} e^{M^2/2\sigma^2} \frac{\sigma}{M}$$

which is the desired formula.

5. INTEGRAL TESTS FOR THRESHOLDS WHICH ARE ALMOST CERTAINLY CROSSED

A further comparison between the jump and Gauss-Markov processes is provided by finding characterizations of those signal excess histories M which decrease so slowly that

(8) $$P(M(t) + \xi_J(t) \geq 0 \qquad \text{for some } 0 \leq t < \infty) = 1$$

or

(9) $$P(M(t) + \xi_G(t) \geq 0 \qquad \text{for some } 0 \leq t < \infty) = 1.$$

Clearly, setting $f = -M$, these problems are equivalent to finding necessary and sufficient conditions on a function f such that

(10) $$P(\xi_J(t) \geq f(t) \qquad \text{for some } 0 \leq t < \infty) = 1$$

or

(11) $$P(\xi_G(t) \geq f(t) \qquad \text{for some } 0 \leq t < \infty) = 1,$$

and we will use this formulation. In all that follows ξ_J is a jump process with parameters λ, σ and ξ_G is a Gauss-Markov process with parameters λ, σ as defined in section 2, $\lambda > 0$, $\sigma \neq 0$.

THEOREM 1: Let f be a positive *nondecreasing* function defined on $[0, \infty)$. Then (10) is satisfied by f if and only if

$$\int_0^\infty \frac{1}{f(t)} e^{-f(t)^2/2\sigma^2} dt = \infty. \tag{12}$$

LEMMA 1: For a nondecreasing function f defined on $[0, \infty)$ we have for any $x \geq 0$

$$H(x) \equiv P(\xi_J(t) \geq f(t) \quad \text{for some } 0 \leq t \leq x) \tag{13}$$

$$= 1 - \Phi(f(0)/\sigma) \exp\left(-\lambda \int_0^x (1 - \Phi(f(t)/\sigma))dt\right)$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy$.

PROOF: More general results of this sort and their proofs are in [2], [4], and [8]. We include a proof of this simple case here for completeness. Let $\tau = \inf\{t : \xi_J(t) \geq f(t)\}$. Then for any $\Delta > 0$,

$$H(x+\Delta) - H(x) = P(x < \tau \leq x + \Delta)$$

$$= P(\tau > x) P(\xi_J(t) > f(t) \quad \text{for } x < t \leq x + \Delta | \tau > x)$$

$$= (1 - H(x)) \{P(\xi_J(t) > f(t) \quad \text{for } x < t \leq x + \Delta \text{ and no jump occurs on this interval } | \tau > x)$$

$$+ P(\xi_J(t) > f(t) \quad \text{for } x < t \leq x + \Delta \text{ and at least one jump occurs on this interval } | \tau > x)\}$$

$$= (1 - H(x)) \{0 + \lambda\Delta(1 - \Phi(f(x + o(\Delta))/\sigma)) + o(\Delta)\}$$

and therefore

$$\lim_{\Delta \to 0} \frac{H(x+\Delta) - H(x)}{\Delta} = \lambda(1 - \Phi(f(x)/\sigma))(1 - H(x)).$$

A similar probability estimate for the case $\Delta < 0$ results and we have the differential equation

$$H'(x) = \lambda(1 - \Phi(f(x)/\sigma))(1 - H(x))$$

with initial condition

$$H(0) = 1 - \Phi(f(0)/\sigma).$$

It is apparent that (13) is the unique solution of this d.e.

PROOF OF THEOREM 1. Using (13) we see that (10) is satisfied if and only if

$$\int_0^\infty (1-\Phi(f(t)/\sigma))dt=\infty. \tag{14}$$

To proceed we use the well known [6, p. 166] double inequality

$$(2\pi)^{-1/2}e^{-1/2x^2}\left(\frac{1}{x}-\frac{1}{x^3}\right)<1-\Phi(x)<(2\pi)^{-1/2}e^{-1/2x^2}\frac{1}{x}, \qquad \text{for all } x>0,$$

and hence

$$(2\pi)^{-1/2}e^{-1/2x^2}\frac{3}{4x}<1-\Phi(x)<(2\pi)^{-1/2}e^{-1/2x^2}\frac{1}{x}, \qquad \text{for all } x>2. \tag{15}$$

Now $f(t)$ is nondecreasing and the rest of the argument breaks up into two cases. If f is bounded above, then clearly (12) and (14) are both satisfied. But, if f is not bounded above, then there exists a t_0 such that $f(t)/\sigma>2$ for all $t>t_0$. Thus (15) implies that

$$\int_{t_0}^\infty (1-\Phi(f(t)/\sigma))dt$$

is finite or infinite according as

$$\int_{t_0}^\infty \frac{1}{f(t)}e^{-f(t)^2/2\sigma^2}dt$$

is finite or infinite, respectively. Thus (12) and (14) are equivalent.

THEOREM 2: Let f be a continuous, increasing and positive function defined on $[0, \infty)$ and such that $e^{-\lambda s}f(s)$ is decreasing. Then (11) is satisfied if and only if

$$\int_0^\infty f(t)e^{-f(t)^2/2\sigma^2}dt=\infty. \tag{16}$$

The following two lemmas are well known and will be used in the proof.

LEMMA 2: [5]. If ξ_G is a Gauss-Markov process on $(-\infty, \infty)$ with parameters λ, σ, then $\zeta(t)=t^{1/2}\xi_G\left(\frac{1}{2\lambda}\ln t\right)$, $t>0$, $\zeta(0)=0$, defines a standard Brownian motion process on $[0, \infty)$ with variance parameter σ^2.

LEMMA 3: (Kolmogorov's test, see [7], p. 36) If ζ is a standard Brownian motion with variance σ^2 and h is a continuous positive function on $[1, \infty)$ such that $t^{-1}h(t)$ is decreasing and $t^{-1/2}h(t)$ is increasing, then

$$P(\zeta(t)\geq h(t) \qquad \text{for some } 1\leq t<\infty)=1$$

if and only if

$$\int_1^\infty t^{-3/2}h(t)e^{-h(t)^2/2t\sigma^2}dt=\infty.$$

PROOF OF THEOREM 2: Let f be a continuous, increasing and positive function defined on $[0, \infty)$ and such that $e^{-\lambda s}f(s)$ is decreasing. Let h be defined by the equation $h(t) = t^{1/2}f\left(\frac{1}{2\lambda}\ln t\right)$, $t < 1$; then h is a continuous positive function on $[1, \infty)$ with $t^{-1}h(t)$ decreasing and $t^{-1/2}h(t)$ increasing. Therefore,

$$P(\xi_G(t) \geq f(t) \quad \text{for some } 0 \leq t < \infty) = P(\zeta(t) \geq h(t) \quad \text{for some } 1 \leq t < \infty) = 1$$

if and only if

$$\infty = \int_1^\infty t^{-3/2}h(t)e^{-h(t)^2/2T\sigma^2}dt$$

$$= \int_1^\infty t^{-1}f\left(\frac{1}{2\lambda}\ln t\right)e^{-f(1/2\lambda \ln t)^2/2\sigma^2}dt$$

$$= \int_0^\infty e^{-2\lambda x}f(x)\,e^{-f(x)^2/2\sigma^2}2\lambda e^{2\lambda x}dx$$

$$= 2\lambda\int_0^\infty f(x)e^{-f(x)^2/2\sigma^2}dx$$

and hence (11) if and only if (16). This proves the theorem.

It is interesting to note a parallel between the formulas (4), (5), and the conditions (12), (16). The formulas (4), (5) each contain a factor e^{M^2/σ^2}; the integrands in (12), (16) each contain a factor $e^{-f^2/2\sigma^2}$; (4) and (5) are identical except for an inversion of a multiplicative factor, while (12) and (16) are identical except for a similar inversion.

It is also of interest to construct a function f which satisfies (16) but not (12). Taking

$$f(t) = \sigma(2\ln t + 3\ln\ln t)^{1/2}, \qquad t > 1, \tag{17}$$

then f satisfies all the regularity conditions of Theorems 1 and 2,

$$\frac{1}{f(t)}e^{-f(t)^2/2\sigma^2} \sim (\sigma\sqrt{2}\,t(\ln t)^2)^{-1}$$

and

$$f(t)e^{-f(t)^2/2\sigma^2} \sim \sigma\sqrt{2}\,(t\ln t)^{-1},$$

so that (16) and hence (11) are satisfied, but (12) and hence (10) are not satisfied, for this threshold f.

Note that for the class of functions f considered, if (12) is satisfied then so is (16). Thus it is easier to have CDP = 1 with the Gauss-Markov process than with the jump process with the same parameters.

Reverting to the original model (1), suppose $M(t) = a - f(t)$ where $f(t)$ is the transmission loss between target and receiver at time t. If f is of the form (17) then detection is sure to occur if the

Gauss-Markov process is used as the model for acoustic fluctuations, but not sure to occur if the jump process is used. This would be a serious conceptual problem if transmission loss behaved like $\sqrt{\ln R}$ where R is range, but fortunately it behaves more like $\ln R$ so this example is probably more of theoretical than practical interest.

REFERENCES

[1] Abramowitz, M. and I. A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (National Bureau of Standards Applied Mathematics Series, U.S. Department of Commerce, Washington, D.C., 1964).

[2] Belkin, B., "First Passage to a General Threshold for a Process Corresponding to Sampling at Poisson Times," Journal of Applied Probability, *8*, 573–588 (1971).

[3] Belkin, B. and M. A. Snyder, "Expansion of the Weber Function D_ν for Small Order with an Application," Journal of Mathematical Analysis and Applications, *38*, 320–327 (1972).

[4] Curtis, E. C., H. R. Richardson, and E. P. Loane, "Cumulative Detection Probability for Continuous-Parameter Stochastic Processes," Daniel H. Wagner, Associated Report to U.S. Naval Air Development Center (1966) (DDC No. AD 803585).

[5] Doob, J. L., "The Brownian Movement and Stochastic Equations," Annals of Mathematics, *43*, 351–369 (1942).

[6] Feller, W., *An Introduction to Probability Theory and its Applications*, Volume I, 2nd Edition (Wiley, New York, 1957).

[7] Ito, K., and H. P. McKean, *Diffusion Processes and Their Sample Paths* (Academic Press, New York, 1965).

[8] Kettelle, J. D. and D. H. Wagner, "Influence of Noise on SSK Effectiveness," Kettelle & Wagner Report to David Taylor Model Basin (1960), Confidential.

[9] Koopman, B. O., unpublished working memorandum to Commander Submarine Development Group Two, Groton, Connecticut, 1970.

[10] Rosser, J. B., Theory and Applications of $\int_0^z e^{-p^2y^2}dy \int_0^y e^{-x^2}dx$. (Mapleton House, Brooklyn, New York, 1948).

[11] Urick, R. J., *Principles of Underwater Sound for Engineers* (McGraw-Hill, New York, 1967).

[12] Commander Submarine Forces Atlantic Instruction S3510.24, "Submarine ASW Search Manual," to appear, Secret.

TARGET SELECTION IN LANCHESTER COMBAT: HETEROGENEOUS FORCES AND TIME-DEPENDENT ATTRITION-RATE COEFFICIENTS

James G. Taylor*

Department of Operations Research and Administrative Sciences
Naval Postgraduate School
Monterey, California

ABSTRACT

We develop solutions to two fire distribution problems for a homogeneous force in Lanchester combat against heterogeneous enemy forces. The combat continues over a period of time with a choice of tactics available to the homogeneous force and subject to change with time. In these idealized combat situations the lethality of each force's fire (as expressed by the Lanchester attrition-rate coefficient) depends upon time. Optimal fire distribution rules are developed through the combination of Lanchester-type equations for combat attrition and deterministic optimal control theory (Pontryagin maximum principle). Additionally, the theory of state variable inequality constraints is used to treat the nonnegativity of force levels. The synthesis of optimal fire distribution policies was facilitated by exploiting special mathematical structures in these problems.

1. INTRODUCTION

Many years ago Karl von Clausewitz said that if theory caused a more critical study of war, then it had achieved its purpose (see p. 191 of [8]). In a companion paper [30], we have studied the structure of the optimal fire distribution policies for some basic "elementary" tactical situations described by Lanchester-type equations of warfare. We did this by contrasting the "best" fire distribution policies for a sequence of scenarios: prescribed-duration battle, terminal-control battle, two enemy target types, many enemy target types, etc. In this companion paper we gave results for some special cases of time dependent attrition-rate coefficients and for reasons of brevity omitted their development. In the paper at hand we present the justification of these results and present extensions: we consider more target types and also replacements for all combatant types.

Thus, the purpose of this paper is to justify some results given in [30] and to extend our past results [26], [30] on optimal fire distribution for a homogeneous force in Lanchester combat against heterogeneous enemy forces. This is done by studying two specific problems. Both battle scenarios that we consider here are for a prescribed duration of time. Additionally, all Lanchester attrition-rate coefficients may change over time. We justify the results given in [30] for a two-on-one combat problem (denoted as (Problem 4) in [30]) by showing that they are a special case of a more general problem considered here.

*This research was supported by the Office of Naval Research as part of the Foundation Research Program at the Naval Postgraduate School.

One direction of generalization of our past results [26], [30] is to consider more target types in cases with variable attrition-rate coefficients: we study the problem of optimal fire distribution for a homogeneous force in Lanchester combat against n different enemy force types for some special cases of variable attrition-rate coefficients. Another direction of generalization is to consider the the nonnegativity restrictions on force levels when replacements are allowed. In this latter case the theory of state variable inequality constraints (see [14]) is employed to develop our results. An important aspect of this is the use of Gamkrelidze's multiplier condition (see Chap. 6 of [21]), which, as this author has pointed out [27], apparently has been overlooked previously in the applications literature (engineering, operations research).

In this paper we consider primarily the application of the mathematical theory of optimal control to develop the optimal fire distribution policy for two problems of interest. Although we do give some interpretation of these results, the interested reader can find a more extensive discussion of the structure of optimal allocation policies and model implications in the referenced companion paper.

In this paper we develop the solutions to all problems by the mathematical theory of optimal control. For the problems at hand we need only make use of the Pontryagin maximum principle [21], although this must be modified when a trajectory lies on the boundary of the state space (which is a compact set) (see Chap. 6 in [21], [10], [11], [12], [20], and [22]). Moreover, the reader should note that there is a sign difference between developments in this country (see p. 108 of [7] or pp. 12–14 of [9]) and those in the Soviet Union [2], [21], although both approaches yield exactly the same results in the applications.

For the first problem that we consider the synthesis of the optimal control is facilitated by taking advantage of the problem's special structure. By the synthesis of the optimal control, we mean the explicit determination of the time history of the optimal control from initial to terminal time. For the problem at hand, this is accomplished by determining a control law by the maximum principle and then working backwards from the end of the problem by a backwards integration of the adjoint system of differential equations for the dual variables which have a boundary condition there.

The organization of this paper is as follows. First, we consider the problem of determining the optimal distribution of fire over n target types by a homogeneous force in Lanchester combat for some special cases of variable attrition-rate coefficients. The optimal policy in a special case is developed for this general problem. Then, particular cases of these general results are related to previous results of this author (in one case justifying the results given for (Problem 4) in [30]), and the structure of the optimal policy is discussed. Second, we consider the problem of determining the optimal distribution of fire over two target types by a homogeneous force in Lanchester combat when replacements are allowed for all force types. Again, we develop our results for some special cases of variable attrition-rate coefficients. These results are developed via the theory of state variable inequality constraints (frequently denoted, for convenience, as SVIC's) and then are related to our previous work on the Isbell and Marlow fire programming problem [13] in which another approach was used. Finally, we make some observations on the results presented in this paper.

2. NOTATION

The symbols which are used in this paper are defined as follows:

a_1, a_2, b_1, b_2 = constant attrition-rate coefficients,

$a_1(t), \ldots, a_n(t), b_1(t), \ldots, b_n(t)$ = variable attrition-rate coefficients,

$c_i(t)$ for $i=1, \ldots, n$ = coefficient of ϕ_i in maximization problem (defined by Equation (16)).
c = constant of proportionality,
$e_i(\tau)$ for $i=1, \ldots, n$ = coefficient of ϕ_i in maximization problem (defined by Equation (27)).
$h(t)$ = variable portion of variable attrition-rate coefficient, e.g. $a_1(t)=k_{a_1}h(t)$,
$h_1(\tau)=h(T-\tau)$,
H = Hamiltonian function,
$k_{a_1}, \ldots, k_{a_n}, k_{b_1}, \ldots, k_{b_n}$ = constant portions of variable attrition-rate coefficients, e.g., $a_1(t)=k_{a_1}h(t)$,
n = number of X-force target types,
p, q, r = utilities assigned per unit of surviving X_1, X_2, and Y forces, respectively,
$p_i(t)$ for $i=1, \ldots, n+1$ = dual variable corresponding to $x_i(t)$ $(x_{n+1}(t)=y(t))$,
$r_1(t), r_2(t), s(t)$ = rate of replacement,
R_i for $i=1, \ldots, n-1 = k_{a_i}(k_{b_i}w_n - k_{b_n}w_i)/(k_{a_i}k_{b_i} - k_{a_n}k_{b_n})$,
S_i for $i=1, \ldots, n = k_{a_i}(k_{b_i}w_k - k_{b_k}w_i)/(k_{a_i}k_{b_i} - k_{a_k}k_{b_k})$,
$\qquad i\neq k$
t = time after beginning of battle,
t_c = time of occurrence of "corner,"
t_e = time of entry to constrained subarc,
$t_{\bar{e}} = \lim_{\substack{t\to t_e \\ t\leq t_e}} t$,
t_l = time of exit (leaving) from constrained subarc,
$t_1 = T-\tau_1$ = time of first switch in extremal tactics,
T = total time for the battle,
T_1 = maximum possible duration for battle, i.e. $T \leq T_1$,

$$u=\int_0^{\tau} b_n(T-\tau)\,d\tau,$$

$V_k = \sqrt{(R_k^2 - w_n^2)k_{a_n}/k_{b_n} + v^2}$,
$w_1, \ldots, w_n, v$ = utilities assigned per unit of surviving $X_1, \ldots, X_n, Y$ forces, respectively,
W = Bellman's optimal value function,
$W_k = k_{a_n}(k_{b_k}w_n - k_{b_n}w_k)/(k_{a_k}k_{b_k} - k_{a_n}k_{b_n})$ where $k \neq n$,
$x_1, \ldots, x_n, y$ = average force strengths; with initial values $x_1^0, \ldots, x_n^0, y_0$,

$$\gamma(\tau, \tau_1) = \sqrt{k_{a_k}k_{b_k}}\int_{\tau_1}^{\tau} h_1(\tau)\,d\tau,$$

$$\delta_{ij} = \text{Kronecker delta} = \begin{cases} 1 & \text{for } i=j, \\ 0 & \text{otherwise,} \end{cases}$$

$\mu_i(t)$ for $i=1, 2,$ = multiplier corresponding to state variable inequality constraint $x_i \geq 0$,
ϕ = fraction of Y-fire directed at X_1,
ϕ_i for $i=1, \ldots, n$ = fraction of Y-fire directed at X_i,
ϕ^* = extremal (sometimes optimal) control,

$$\theta(\tau) = \sqrt{k_{a_n}k_{b_n}}\int_0^{\tau} h_1(\tau)\,d\tau,$$

τ = "backwards time" from the end of the battle; defined by $\tau = T - t$, i.e. the time remaining before the end of battle,

τ_1, τ_2, etc. = "backwards time" of the first, second, etc., switch in extremal tactics.

3. SEVERAL TARGET TYPES, SOME SPECIAL CASES OF VARIABLE ATTRITION-RATE COEFFICIENTS

We consider the following prescribed duration battle against n target types:

$$\underset{\phi_i(t)}{\text{maximize}} \left\{ vy(T) - \sum_{i=1}^{n} w_i x_i(T) \right\} \qquad \text{with } T_1 \text{ specified,}$$

subject to:

$$\frac{dx_i}{dt} = -\phi_i a_i(t) y \qquad \text{for } i = 1, \ldots, n,$$

$$\frac{dy}{dt} = -\sum_{i=1}^{n} b_i(t) x_i,$$

(1)

$$x_i, y \geq 0, \qquad \phi_i \geq 0 \qquad \text{for } i = 1, \ldots, n, \qquad \sum_{i=1}^{n} \phi_i = 1, \qquad \text{and } T \leq T_1,$$

where all symbols are defined in Section 2 above.

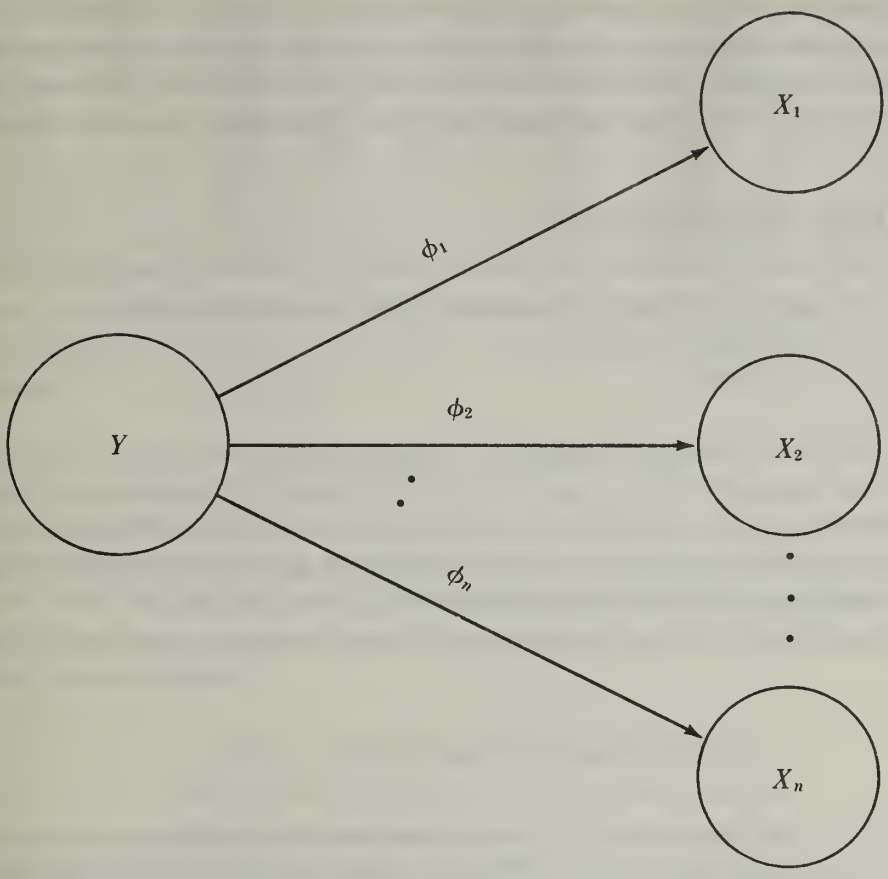
In the above problem (1) $x_1, \ldots, x_n$, and y are called *state* variables, while $\phi_1, \ldots, \phi_n$ are called *control* (or decision) variables. A constraint such as $x_1 \geq 0$ is called a state variable inequality constraint (SVIC) and requires special treatment (see Chap. 6 of [21]). In other words, the well-known maximum principle (as presented in Chaps. 1–3 of [21]*) requires modification in problems with SVIC's (see also Chap. 3 of [7]). However, to avoid inessential complications only the special case in which $x_i(t) > 0$ for $i = 1, \ldots, n$ and $y(t) > 0$ for $0 \leq t < T$ is considered for our developments of Sec. 3. In Sec. 4, we use the theory of SVIC's to treat explicitly the nonnegativity of force levels for a version of the above problem (1) in which replacements are considered.

The physical interpretation of this idealized military situation is as follows: combat for a known length of time between an X force composed of n types of weapon systems and a homogeneous Y force. The objective of the Y-force commander is to maximize the military worth of his surviving forces at the end of a battle lasting a known length of time T and to minimize that of the enemy's. This is accomplished by his choice of the fraction of fire, ϕ_i, directed at the X_i forces. A schematic of this scenario is shown below. It is assumed that a military worth can be assigned to survivors of each force-type and that this utility is additive. We denote this utility per unit of weapon system as $w_1, \ldots, w_n$, and v for the $X_1, \ldots, X_n$, and Y forces, respectively.

Our state equations, Equation (1), i.e., the differential equation constraints, represent the (deterministic) attrition process in this idealized battle. We shall refer to attrition as being a "square-law" process when the casualty rate is proportional to only the number of enemy firers and as being a "linear-law" process when it is proportional to the product of the numbers of enemy firers and remaining targets. For the problems studied in this paper we assume a "square-law" process. The interested reader can find discussions of the physical assumptions which lead to this type of attrition in articles by H. Weiss [32] and H. Brackney [6]. The essential points are that new targets are acquired at a rate independent of force levels and that fire is aimed at point targets.

---

*In this paper we employ an equivalent version commonly used in the control theory literature of this country (see pp. 108–109 of [7] and pp. 12–14 of [9]).



S. Bonder's article [4] (published in 1967 (see also [3])) is the pioneering work on the determination of the Lanchester attrition-rate coefficient (for weapon systems that adjust fire based on the results of the immediately preceding round). Subsequent work by Barfoot [1], Bonder [5], and Kimbleton [17] has led to the development of formulas for estimating the Lanchester attrition-rate coefficient, the rate at which one unit of weapon system destroys enemy targets, as a synthesis of the following factors: hit probabilities, rates of fire, target acquisition rate, weapon system projectile-target lethality characteristics, adjustment process. Other papers [23] and [24] have discussed the statistical estimation of parameters for the Lanchester attrition-rate coefficient from empirical data.

In general, the Lanchester attrition-rate coefficient is range dependent, and the above formulation (1) takes this into account by having the attrition-rate coefficients depend upon time. S. Bonder [3] has done the pioneering work on analyzing dynamic combat situations with Lanchester-type equations and variable attrition rates (also see [25], [28], and [29] for more background). For example, consider a mobile attack on a static defensive position. The effectiveness (Lanchester attrition-rate coefficient) of weapons systems depends upon force separation, and this is, in turn, related to time via the attack velocity. In [25] we showed how to develop Lanchester-type equations with either time or range as the independent variable and noted the equivalence of the two formulations (see [29] for more extensive results with variable attrition-rate coefficients). Hence, the above was our motivation for studying scenarios with time-dependent attrition rates.

We finally note that when fire is uniformly distributed over a target area [32] or the rate of target destruction is constrained by the rate of acquisition of new targets and this is inversely proportional to target density [6], a "linear-law" attrition process results. We have elsewhere [30] pointed out that this type of attrition structure leads to a fundamentally different structure for the optimal fire distribution policies.

3.1. Optimal Policy in a Special Case

The battle lasts for $0 \leq t \leq T_1$ unless, of course, one side or the other is annihilated before T_1. To be more precise, the battle terminates under one of the following three conditions:

(1) $x_1(T) = \ldots = x_n(T) = 0$ and $T \leq T_1$,

(2) $y(T) = 0$ and $T \leq T_1$,

(3) $T = T_1$,

where T denotes the time at which the battle ends. However, for our developments of this section, only the special case in which $x_1(t) > 0, \ldots x_n(T) < 0, y(T) > 0$, and $T = T_1$ is considered. In other words, those subcases in which a state variable is reduced to zero are not considered. Thus, it is assumed that the initial force levels are such that no force type is annihilated during this prescribed duration battle. Explicit consideration is given to the nonnegativity of force levels in Sec. 4.

In all cases, we assume that the X-force weapon systems have performance characteristics, such that

$$b_i(t) = k_{bi} h(t) \qquad \text{for } i = 1, \ldots, n. \tag{2}$$

This assumption physically means that all X-force weapon systems have the same type of range capability (for example, quadratic dependence of kill rate on range) and the same effective range, although one weapon system dominates the others in exactly the same manner at all ranges.

It seems appropriate to point out that there is a special instance in which the optimal allocation policy takes a particularly simple form. (We first made this type of observation in [30] (see also [31]).) This is when the Y force values surviving X-force types in direct proportion to their kill rate against the Y force at the end of battle, i.e.,

$$w_i = c b_i(t = T) = c k_{bi} h(t = T). \tag{3}$$

In this case, we have

$$\frac{w_i}{w_j} = \frac{k_{bi}}{k_{bj}}. \tag{4}$$

Assuming (3), the optimal fire distribution rule consequently takes a particularly simple form for $0 \leq t \leq T$

$$\phi_i^*(t) = \delta_{i, j(t)}, \tag{5}$$

where $\delta_{i,j}$ denotes the Kronecker delta and is equal to 1 for $i = j$ and zero otherwise and $j(t)$ is the index such that

$$a_j(t) b_j(t) = \text{maximum } (a_1(t) b_1(t), \ldots, a_n(t) b_n(t)). \tag{6}$$

In this case the optimal fire distribution policy depends only upon the product of attrition-rate coefficients which may be interpreted as the rate of destruction of enemy kill-rate capability. All fire is concentrated on the target type with the largest product of attrition-rate coefficients.* Target priority is subject to change over time as the ranking of target types according to this decision criterion changes. It is possible for the optimal tactic to be to shift fire from one target type to another several times over the course of battle with the duration of battle not having any direct effect upon this. Observe that no assumptions at all have been made about the Y-force attrition-rate coefficients against $X_1, \ldots,$ and X_n, i.e., $a_1(t), \ldots,$ and $a_n(t)$.

We now consider the more complex case in which enemy force types are, in general, not valued in proportion to their kill rates. Again, we consider a special case when relatively simple analytic results are still possible. Thus, we assume that

$$a_i(t) = k_{a_i} h(t) \qquad \text{for } i = 1, \ldots, n. \tag{7}$$

In the case, for example, of a mobile attack against a static defensive position, assumptions (2) and (7) have the physical interpretation that all weapon systems and weapon system combinations have the same effective range and the same type of range dependency for their kill rate against any target type. The solution for this case is shown in Table I. These results are discussed further below. We now, however, consider their development.

TABLE I. *Solution to fire distribution problem*

Battle of prescribed duration with variable attrition-rate coefficients

Special case in which $x_i(T) > 0$ for $i = 1, \ldots, n$ and $y(T) > 0$

Special assumption: $a_i(t) = k_{a_i} h(t)$ and $b_i(t) = k_{b_i} h(t)$ for $i = 1, \ldots, n$.

Case A: $w_i = cb_i(t = T)$ for $i = 1, \ldots, n$

Optimal control

$\phi_i^*(t) = \delta_{i,j(t)}$ for $0 \leq t \leq T$

$i = 1, \ldots, n$

Case B: $w_i \neq cb_i(t = T)$ for at least one index i

Optimal control

(a) for $\tau_1 \geq T$

$\phi_i^*(t) = \delta_{in}$ for $0 \leq t \leq T$

$i = 1, \ldots, n$

(b) for $\tau_2 \geq T > \tau_1$

$$\phi_i^*(t) = \begin{cases} \delta_{ik} & \text{for } 0 \leq t \leq T - \tau_1 \\ \delta_{in} & \text{for } t - \tau_1 < t \leq T \end{cases}$$

$i = 1, \ldots, n$

(c) for $\tau_3 \geq T > \tau_2$

$$\phi^*(t) = \begin{cases} \delta_{ij} & \text{for } 0 \leq t < T - \tau_2 \\ \delta_{ik} & \text{for } T - \tau_2 < t \leq T - \tau_1 \\ \delta_{in} & \text{for } T - \tau_1 < t \leq T \end{cases}$$

$l = 1, \ldots, n$

etc.

See Notes to Table 1 on page 390.

*One must, of course, also show the impossibility of a singular solution [15], [16] with a singular control other than the 0–1 optimal allocation given above (5) (see [31]). This is done later in the paper at hand.

NOTES:

(1) $J(t)$ is index such that $a_J(t)b_J(t) = \max(a_1(t)b_1(t), \ldots, a_n(t)b_n(t))$.

(2) n is index assigned so that $a_n(t=T)w_n = \max(a_1(t=T)w_1, \ldots, a_n(t=T)w_n)$.

(3) k is index such that $R_K = \min_{\substack{R_i>0 \\ k_{a_i}k_{b_i} > k_{a_n}k_{b_n}}} (R_1, \ldots, R_{n-1})$ where $R_i = \dfrac{k_{a_i}(k_{b_i}w_n - k_{b_n}w_i)}{k_{a_i}k_{b_i} - k_{a_n}k_{b_n}}$ for $i = 1, \ldots, n-1$.

(4) τ_1 is given by $\displaystyle\int_{T-\tau_1}^{T} h(t)dt = \frac{1}{\sqrt{k_{a_n}k_{b_n}}} \ln \left\{ \frac{\left(\frac{R_k}{w_n}\right)^2 + \sqrt{\left(\frac{R_k}{w_n}\right)^2 + \left(\frac{v}{w_n}\right)^2 \frac{k_{b_n}}{k_{a_n}} - 1}}{1 + \frac{v}{w_n}\sqrt{\frac{k_{a_n}}{k_{a_n}}}} \right\}$

(5) j is index such that $S_j = \min_{\substack{S_i>0 \\ k_{a_i}k_{b_i} > k_{a_k}k_{b_k} \\ i \neq k}} (S_1, \ldots, S_n)$ where $S_i = \dfrac{k_{a_i}(k_{b_i}w_k - k_{b_k}w_i)}{k_{a_i}k_{b_i} - k_{a_k}k_{b_k}}$ for $i = 1, \ldots, n$, $i \neq k$.

(6) τ_2 is given by $\displaystyle\int_{T-\tau_2}^{T-\tau_1} h(t)dt = \frac{1}{\sqrt{k_{a_k}k_{b_k}}} \ln \left\{ \frac{\left(\frac{S_j}{W_k}\right)^2 + \sqrt{\left(\frac{S_j}{W_k}\right)^2 + \left(\frac{V_k}{W_k}\right)^2 \frac{k_{b_k}}{k_{a_k}} - 1}}{1 + \frac{V_k}{W_k}\sqrt{\frac{k_{b_k}}{k_{a_k}}}} \right\}$

(7) τ_3 is given by expression similar to those for τ_1 and τ_2 above.

3.2. Development of Extremal Fire Distribution Policy

For $y > 0$ and $x_1 > 0$ for $i = 1, \ldots, n$, the Hamiltonian for (1) is given by [7]

$$H(t, x_i, p_i, \phi_i) = -y \sum_{i=1}^{n} a_i(t)p_i(t)\phi_i - p_{n+1}(t) \sum_{i=1}^{n} b_i(t)x_i, \tag{8}$$

where $p_i(t)$ for $i = 1, \ldots, n$ denotes the dual variable corresponding to x_i and $p_{n+1}(t)$ denotes the dual variable corresponding to y.* According to the maximum principle, the nonsingular optimal control (when there is only one extremal (see Sec. 3.3 below)) is determined by the (trivial) linear program

$$\underset{\phi_i}{\text{maximize}}\ H(t, x_i, p_i, \phi_i),$$

subject to:

$$\sum_{i=1}^{n} \phi_i = 1,$$

$$\phi_i \geq 0,$$

which, in turn, leads to

$$\underset{\phi_i}{\text{maximize}} \sum_{i=1}^{n} a_i(t)(-p_i(t))\phi_i, \tag{9}$$

---

*The reader should recall the well-known interpretation of a dual variable $p_i(t)$ as the effect on the total value of survivors at the end of battle when an optimal policy is followed from time t until time T of having an additional unit of $x_i(t)$ [here we have let $y(t) = x_{n+1}(t)$]. Another way to state this is that

$$p_i(t) = \frac{\partial W}{\partial x_i(t)} \qquad \text{where } W = W(t, x_1, \ldots, x_n, y)$$

denotes Bellman's optimal value function.

subject to:
$$\sum_{i=1}^{n} \phi_i = 1,$$

$$\phi_i \geq 0.$$

By inspection the solution to (9) is easily seen to be

(10) $$\phi_i^*(t) = \delta_{i,j(t)},$$

where $j(t)$ is the index such that

$$a_j(t)(-p_j(t)) = \text{maximum } (a_1(t)(-p(t)), a_2(t)(-p(t)), \ldots, a_n(t)(-p_n(t))).$$

To trace the history of ϕ_i^* over time, we consider the adjoint system of differential equations given by

(11) $$\frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i} = b_i(t)p_{n+1} \text{ with } p_i(t=T) = -w_i \qquad \text{for } i=1, \ldots, n,$$

and

$$\frac{dp_{n+1}}{dt} = -\frac{\partial H}{\partial y} = \sum_{i=1}^{n} \phi_i^* a_i(t) p_i \text{ with } p_{n+1}(t=T) = v.$$

Additionally, at a corner which occurs at an interior point of the state space (i.e., $x_i > 0$ for $i=1, \ldots, n$ and $y > 0$) the following well-known corner conditions hold [7] (see also [31])

(12) $$p_i(t_c^-) = p_i(t_c^+) \qquad \text{for } i=1, \ldots, n+1,$$

and

(13) $$H^*(t_c^-) = H^*(t_c^+),$$

where t_c^- denotes the time just before the corner, t_c^+ denotes the time just after the corner, and $H^*(t_c^-)$ denotes $H(t=t_c^-, x_i^*, p_i, \phi_i^*)$. Such a corner occurs at the "switching times" for the "bang-bang" control law (10).

From the assumption (2) and the adjoint equations, Equation (11), it follows that

$$\frac{dp_i}{dp_n} = \frac{k_{b_i}}{k_{b_n}},$$

which, when integrated, leads to

(14) $$p_i(t) = \frac{k_{b_i}}{k_{b_n}} \{p_n(t) + w_n\} - w_i \qquad \text{for } i=1, \ldots, n.$$

Using (14), we may write (9) as

$$\underset{\phi i}{\text{maximize}} \sum_{i=1}^{n} \phi_i c_i(t), \tag{15}$$

subject to:

$$\sum_{i=1}^{n} \phi_i = 1,$$

where

$$\phi_i \geq 0,$$

$$c_i(t) = \frac{(-p_n(t))}{k_{bn}} a_i(t) k_{bi} \left[1 + \frac{w_i}{(-p_n(t))} \left\{\frac{k_{bn}}{k_{bi}} - \frac{w_n}{w_i}\right\}\right]. \tag{16}$$

The special assumption (3) that enemy force types are valued in direct proportion to their kill rate against Y at the end of battle leads to simplification of (16), namely

$$c_i(t) = \frac{(-p_n(t))}{k_{bn}} a_i(t) k_{bi} = \frac{(-p_n(t))}{b_n(t)} a_i(t) b_i(t), \tag{17}$$

whence our result (5), since it is readily shown that $p_n(t) < 0$.

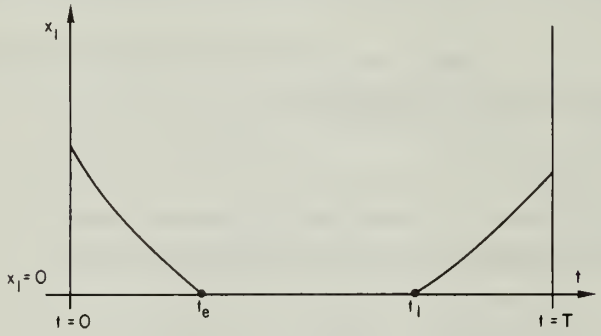


FIGURE 1. Entry to and exit from constrained subarc.

It is now further assumed that (7) holds. The extremal control* shown in Table I is developed as follows. Since we develop the solution by working backwards from the end of battle at $t=T$ (where boundary values are known for the dual variables, i.e., $p_i(t=T)$), it is convenient to introduce the "backwards time" variable τ defined by $\tau = T - t_0$. It is assumed that the enemy target types are indexed so that

$$a_n(t=T)w_n = a_n(t=T)(-p_n(t=T)) = \text{maximum } (a_1(t=T)w_1, \ldots, a_n(t=T)w_n). \tag{18}$$

By (9) it is easily seen that

$$\phi_i^*(\tau) = \delta_{in} \qquad \text{for } \tau \in [0, \tau_1], \tag{19}$$

where τ_1 denotes the "backwards time" of the first switch in the extremal fire distribution policy. Giving consideration to (19) and observing that $\frac{d}{dt} = -\frac{d}{d\tau}$, we see that for $\tau \in [0, \tau_1]$ one need only consider the following equations from the adjoint system (11)

*The optimality of this control is discussed in Sec. 3.3.

$$\frac{dp_n}{d\tau} = -b_n(t)p_{n+1} \qquad \text{with } p_n(\tau=0) = -w_n,$$

(20)

$$\frac{dp_{n+1}}{d\tau} = -a_n(t)p_n \qquad \text{with } p_{n+1}(\tau=0) = v,$$

and it is recalled (rewriting (14)) that

$$p_i(\tau) = \frac{k_{bi}}{k_{bn}}\{p_n(\tau) + w_n\} - w_i \qquad \text{for } i = 1, \ldots, n.$$

The above variable coefficient adjoint differential equations, Equation (20), are readily integrated, however, by consideration of (2) and (7), i.e., $a_n(t)/b_n(t) = \text{constant}$, and the observations elaborated upon in [29], i.e., from (20) we can obtain the following linear second order differential equation for $p_n(\tau)$

$$\frac{d^2p_n}{du^2} - \frac{a_n(t)}{b_n(t)}p_n = 0,$$

with

$$p_n(u=0) = -w_n,$$

$$\frac{dp_n}{du}(u=0) = -v,$$

and

$$u = \int_0^\tau b_n(T-\tau)d\tau.$$

Hence, we readily find that

(22) $$p_n(\tau) = -w_n \cosh \theta(\tau) - v\sqrt{\frac{k_{bn}}{k_{an}}} \sinh \theta(\tau),$$

where

(23) $$\theta(\tau) = \sqrt{k_{an}k_{bn}} \int_0^\tau h_1(\tau)d\tau,$$

and

$$h_1(\tau) = h(T-\tau).$$

It may be that the index $j(t)$ in (10) is not unique, i.e., the linear program (9) has alternate optima. This causes no difficulty unless this situation continues for a finite interval of time. When this happens, the corresponding segment of the battle trajectory is called a *singular subarc* [15], [16]. However, it is easily shown by arguments similar to those given in [30] that it is impossible to have a singular solution to (1). If $j(t)$ were not unique for a finite interval of time, then (for example) one would have $a_i(t)p_i(t) = a_j(t)p_j(t)$ for $t_1 \leq t \leq t_2$. By (7), this becomes

(24) $$k_{a_i}p_i(t) = k_{a_j}p_j(t).$$

If this were to occur, then we must have

$$k_{a_i}\frac{dp_i}{dt}=k_{a_j}\frac{dp_j}{dt},$$

or, using (11) and (2),

$$p_{n+1}(t)h(t)(k_{a_i}k_{b_i}-k_{a_j}k_{b_j})=0. \tag{25}$$

Since $p_{n+1}(t)>0$ and $h(t)>0$ for $0\leq t\leq T$, Equation (25) implies that $k_{a_i}k_{b_i}=k_{a_j}k_{b_j}$ which, in general, is not true. Hence, there is no singular solution to (1), and $\phi_i^*(t)$ is either 0 or 1 (almost everywhere).

We will now determine what conditions lead to a change in the extremal fire distribution policy and the time at which this change occurs, τ_1. To do this it is convenient to rewrite (15) and (16) as

$$\underset{\phi_i}{\text{maximize}}\sum_{i=1}^{n}e_i(\tau)\phi_i, \tag{26}$$

subject to:

$$\sum_{i=1}^{n}\phi_i=1,$$

$$\phi_i\geq 0,$$

where

$$e_i(\tau)=a_i(t)w_i\left[1+\frac{k_{b_i}}{w_ik_{b_n}}\{(-p_n(\tau))-w_n\}\right]. \tag{27}$$

Considering (9), we see that a switch in the extremal distribution of fire occurs at the smallest τ for which

$$a_i(t)w_i\left[1+\frac{k_{b_i}}{w_ik_{b_n}}\{(-p_n(\tau))-w_n\}\right]=a_n(t)(-p_n(\tau)), \tag{28}$$

where i is one of the integers from $\{1, 2, \ldots, n-1\}$ and certain other conditions (to be determined below) are met. Let k denote the index of the target type to which fire is first shifted in "backwards time." Assuming that the index n in (18) is unique, we observe that at $\tau=0$ one has

$$a_i(t=T)w_i<a_n(t=T)w_n \qquad \text{for } i=1, \ldots, n-1. \tag{29}$$

Then for $\tau_1<\tau<\tau_2$, where τ_2 denotes the "backwards time" of the second switch in fire distribution, one has $\phi_i^*(\tau)=\delta_{ik}$, and thus by (7), (26), and (27) the following inequality must hold (using the fact that $h(t)>0$)

$$k_{a_k}w_k\left[1+\frac{k_{b_n}}{w_kk_{b_n}}\{(-p_n(\tau))-w_n\}\right]>k_{a_n}(-p_n(\tau)),$$

which may be re-arranged to yield

$$k_{a_k}(k_{b_k}w_n-k_{b_n}w_k)<(k_{a_k}k_{b_k}-k_{a_n}k_{b_n})(-p_n(\tau)). \tag{30}$$

From (30) it is readily shown by arguments similar to those given in [30] for the constant attrition-rate coefficient case that a necessary condition for fire to be shifted from target type n to target type k when

one works backwards from the end is that

$$k_{a_k}k_{b_k} > k_{a_n}k_{b_n}. \tag{31}$$

Thus, when (31) is true, by (28) with $i = k$ the switch in fire distribution occurs at $\tau = \tau_1$, such that

$$\frac{k_{a_k}(k_{b_k}w_n - k_{b_n}w_k)}{(k_{a_k}k_{b_k} - k_{a_n}k_{b_n})} = (-p_n(\tau = \tau_1)) > 0, \tag{32}$$

where we have used the assumption (7). From (32), we see that a necessary and sufficient condition for fire to be shifted from target type n to target type k when one works backwards from the end is that (31) be true and

$$\frac{k_{b_k}}{w_k} > \frac{k_{b_n}}{w_n}, \tag{33}$$

or, equivalently, by (2)

$$\frac{b_k(t)}{w_k} > \frac{b_n(t)}{w_n}. \tag{34}$$

The above results may be summarized as shown in Table I. The time of switching fire to the kth target type, τ_1, is determined by the equation (combining (7), (22), and (28) with $i = k$)

$$w_n \cosh \theta(\tau = \tau_1) + v\sqrt{\frac{k_{b_n}}{k_{a_n}}} \sinh \theta(\tau = \tau_1) = \frac{k_{a_k}(k_{b_k}w_n - k_{b_n}w_k)}{k_{a_k}k_{b_k} - k_{a_n}k_{b_n}}, \tag{35}$$

where $\theta(\tau) = \sqrt{k_{a_n}k_{b_n}} \int_{T-\tau}^{T} h(t)dt$. Equation (35) may be solved to yield the expression for τ_1 given in Table I.

The general pattern of when and to which target type fire is shifted as one works backwards from the end of battle does not emerge until one has considered the second switch in fire distribution. Since this is dependent upon the backwards evolution of target worth, the backwards integration of the adjoint system of differential equations, Equation (11), must be further considered. From above, one has that

$$\phi_i^*(\tau) = \delta_{ik} \qquad \text{for } \tau \epsilon (\tau_1, \tau_2), \tag{36}$$

where τ_2 denotes the "backwards time" of the second switch in fire distribution. From (36), it is seen that for $\tau \epsilon [\tau_1, \tau_2]$ one needs only to consider the following equations from the adjoint system (11)

$$\frac{dp_k}{d\tau} = -b_k(t)p_{n+1} \qquad \text{with } p_k(\tau = \tau_1) = -W_k,$$

$$\frac{dp_{n+1}}{d\tau} = -a_k(t)p_k \qquad \text{with } p_{n+1}(\tau = \tau_1) = V_k, \tag{37}$$

where

$$W_k = \frac{k_{a_n}(k_{b_k}w_n - k_{b_n}w_k)}{k_{a_k}k_{b_k} - k_{a_n}k_{b_n}}, \tag{38}$$

and

$$V_k = \sqrt{\frac{k_{a_n}}{k_{b_n}}(R_k^2 - w_n^2) + v^2}. \tag{39}$$

Equation (38) follows from the fact that by (9) and (10) at $\tau = \tau_1$ we have

$$a_k(t = T - \tau_1)(-p_k(\tau_1)) = a_n(t = T - \tau_1)(-p_n(\tau_1)), \tag{40}$$

which may be combined with (7) and (32) to yield the desired result. Equation (39) is developed as follows. For $0 \leq \tau \leq \tau_1$, we have

$$\frac{dp_n}{dp_{n+1}} = \frac{k_{b_n} p_{n+1}}{k_{a_n} p_n},$$

which is readily integrated to yield a "square law" between the dual variables $p_n(\tau)$ and $p_{n+1}(\tau)$ for $0 \leq \tau \leq \tau_1$

$$k_{a_n}\{p_n^2(\tau) - w_n^2\} = k_{b_n}\{p_{n+1}^2(\tau) - v^2\}, \tag{41}$$

whence follows (39). Additionally, in developing the above results for $p_n(\tau = \tau_1)$ and $p_{n+1}(\tau = \tau_1)$ we have used the corner conditions (12). It should be noted that all the dual variables may be expressed in terms of $p_k(\tau)$ (let $n = k$ in (14)). Thus, we have

$$p_i(\tau) = \frac{k_{b_i}}{k_{b_k}}\{p_k(\tau) + w_k\} - w_i \qquad \text{for } i = 1, \ldots, n. \tag{42}$$

Again, Equations (37) may be solved to yield for $\tau \epsilon [\tau_1, \tau_2]$

$$p_k(\tau) = -W_k \cosh \gamma(\tau, \tau_1) - V_k \sqrt{\frac{k_{b_k}}{k_{a_k}}} \sinh \gamma(\tau, \tau_1), \tag{43}$$

where

$$\gamma(\tau, \tau_1) = \sqrt{k_{a_k} k_{b_k}} \int_{\tau_1}^{\tau} h_1(\tau) d\tau, \tag{44}$$

and $h_1(\tau) = h(T - \tau)$.

Subsequent arguments are now similar to those given for the first switch in extremal fire distribution. Let j denote the index of the target type to which fire is shifted secondly in "backwards time." Then, it may be shown by arguments similar to the above that necessary conditions for fire to be shifted to the jth target type are that

$$k_{a_j} k_{b_j} > k_{a_k} k_{b_k} > k_{a_n} k_{b_n}, \tag{45}$$

and

$$\frac{k_{b_j}}{w_j} > \frac{k_{b_k}}{w_k} > \frac{k_{b_n}}{w_n}, \tag{46}$$

or, equivalently, by (2) and (7)

$$a_j(t) b_j(t) > a_k(t) b_k(t) > a_n(t) b_n(t), \tag{47}$$

and

$$\frac{b_j(t)}{w_j} > \frac{b_k(t)}{w_k} > \frac{b_n(t)}{w_n} . \tag{48}$$

Using an argument similar to the one for the first switch in fire distribution, it may be shown that the target type to which fire is shifted secondly (working backwards from the end of battle) has index j determined as shown in Table I. The time of switch, τ_2, of fire to the jth target type is determined by the transcendental equation

$$W_k \cosh \gamma(\tau_2, \tau_1) + V_k \sqrt{\frac{k_{b_k}}{k_{a_k}}} \sinh \gamma(\tau_2, \tau_1) = \frac{k_{a_j}(k_{b_j} w_k - k_{b_k} w_j)}{k_{a_j} k_{b_j} - k_{a_k} k_{b_k}} , \tag{49}$$

which may be solved to yield the expression for τ_2 given in Table I. Further shifts in fire follow the pattern established above.

3.3 Determination of the Optimal Control

As is well-known, the maximum principle provides only necessary conditions of optimality. Moreover, let the reader recall that we refer to a path on which the maximum principle is everywhere satisfied as an *extremal.* Thus, in order to determine the optimal allocation policy one must show that a particular extremal is indeed an optimal trajectory.

Two ways of demonstrating the optimality of an extremal trajectory are as follows (see also [31]):

(a) check whether sufficient conditions of optimality are satisfied on the extremal,

(b)* by citing the appropriate existence theorem, show that an optimal control exists to the problem at hand; there are two further subcases: (1) if the extremal is unique, then it is optimal or (2) if the extremal is not unique and only a finite number exist, then the optimal trajectory is determined by considering the finite number of alternatives.

It has not been possible to take the former approach via standard control theory arguments (see pp. 181–184 of [7]), since the problem at hand (1) is singular (in the sense that $\partial^2 H/\partial\phi^2 \equiv 0$). Moreover, the results of Funk and Gilbert [10] do not apply to the problem at hand, since the planning horizon is not of fixed length. This is because the battle can be terminated before the prescribed maximum possible duration T_1 (see Sec. 3.1. above).

Thus, the second approach outlined above is the one that we have taken in this research. Moreover, in the present paper we do not completely carry out this demonstration (as we did for a constant attrition-rate coefficient case in [26]), but we will limit ourselves to a few remarks. If an extremal leads from an initial point in the state space to the terminal surface, then it is optimal and no difficulty exists. We have considered extremals for only the special termination condition $x_1(T), \ldots, x_n(T) > 0$, $y(T) > 0$, and $T = T_1$. When only one extremal leads to this terminal state, it is optimal (provided an extremal path does not lead to any other terminal state from an initial point in the force-level space). However, a thorough examination of the uniqueness of extremals is beyond the scope of the present work.

The existence of an optimal control is established by citing an existence theorem due to Lee and Markus (see Corollary 2 on p. 262 of [18]). This corollary is applicable to the problem (1), since the control variable appears linearly and the responses to the controller ϕ are uniformly bounded. (This latter condition is a consequence of the fact that ϕ is restricted to lie in a compact set and appears

*This essentially is the approach taken by us in [26]. However, we did not cite an existence theorem to guarantee the existence of an optimal control in this previous paper.

linearly in the state equations, Equation (1).) Moreover, the above also establishes the existence of an optimal control for the Isbell and Marlow fire programming problem [26]. Previously, we had implicitly assumed the existence of an optimal control.

3.4. Some Particular Cases of the Optimal Policy of Section 3.1

It seems appropriate to point out that the results developed in this paper for the problem (1) (and primarily given in Table I) are very general and contain many of our previous results [30] as special cases. When $n = 2$, our results of Sec. 3.1. reduce to those given for (Problem 4) of [30]. Thus, we have justified these results previously stated without proof. For constant attrition-rate coefficients the results of Sec. 3.1. reduce to those given for (Problem 3) of [30].

3.5. Discussion of Structure of Optimal Policy

In a companion paper [30] the interested reader can find an extensive discussion of the structure of the optimal fire distribution policies for special cases of the general problem considered in this paper. We have contrasted there the structures of the optimal allocation policies for numerous tactical allocation problems. However, we will further discuss here the optimal fire distribution policy for the problem at hand.

First, let us note that for the problem studied above the optimal control was always "bang-bang," i.e., an extreme point of the constrained control variable space (a compact set). This was proven by showing the impossibility of a singular solution (involving an optimal allocation of a fraction of fire other than 0 or 1). Thus, we saw that the optimal tactic was always to concentrate all fire on a single target type. As we pointed out in [30], this is a consequence* of the "square-law" attrition process (casualty rate proportional to only the number of enemy firers) for X-force target types in (1). Moreover, in this case (at least for the idealized situation studied in this paper) we have a quantitative justification of one of the most significant and oft-quoted of Napoleon Bonaparte's sayings (see p. 117 of [19])—"The principles of war are the same as those of a siege; fire must be concentrated at one point."

Secondly, for the general case of the optimal distribution of fire of a homogeneous force in Lanchester combat against heterogeneous forces (i.e., problem (1)), when enemy survivors were valued in proportion to their effectiveness (i.e., kill rate against the Y forces), the optimal tactic took a remarkably simple form: always concentrate all fire on the target type for which $a_i(t)b_i(t)$ is the largest. The reader should recall that for this problem our basic assumption was that $b_i(t) = k_{b_i} h(t)$ and that we considered only the special case in which $x_1(T) > 0, \ldots, x_n(T) > 0, y(T) > 0$, and $T = T_1$.

Thirdly, to obtain simple analytic results when X-force types were not valued in proportion to their kill rate, we further assumed that $a_i(t) = k_{a_i} h(t)$. Then as the battle developed forwards in time, the concentration of all fire on one target type was always shifted to a new target type for which both quantities $a_i(t)b_i(t)$ and $b_i(t)/w_i$ were smaller than those corresponding to the previously engaged target type. It should be noted, however, that if $k_{a_n} k_{b_n} > k_{a_i} k_{b_i}$ (or, equivalently, $a_n(t)b_n(t) > a_i(t)b_i(t)$) for $i = 1, \ldots, n-1$, then there is no index k satisfying note (3) of Table I. (The reader should recall that the index n is assigned so that $a_n(t = T)w_n > a_i(t = T)w_i$ for $i = 1, \ldots, n-1$.) In this case we have $\tau_1 = +\infty$ so that $\phi_i^*(t) = \delta_{in}$ for all t.

---

* In [31] we show that the optimal allocation policy may be to divide one's fire over two enemy target types when each of the target types undergoes a "linear-law" attrition process.

4. A MODEL WITH REPLACEMENTS: CONSIDERATIONS ON NONNEGATIVITY OF FORCE LEVELS

We consider the following prescribed duration battle against two target types:

$$\underset{\phi(t)}{\text{maximize}} \{ry(T) - px_1(T) - qx_2(T)\} \text{ with } T_1 \text{ specified,}$$

subject to:

$$\frac{dx_1}{dt} = -\phi a_1(t)y + r_1(t),$$

(50)

$$\frac{dx_2}{dt} = -(1-\phi)a_2(t)y + r_2(t),$$

$$\frac{dy}{dt} = -b_1(t)x_1 - b_2(t)x_2 + s(t),$$

$$x_1, x_2, y \geq 0, \quad 0 \leq \phi \leq 1, \quad \text{and } T \leq T_1,$$

where all symbols are defined in Sec. 2. above. The reader should note that replacements are allowed for all force types. In this section we will give explicit consideration to the nonnegativity of the force levels. As noted in Sec. 3. a constraint such as $x_1 \geq 0$ is called a state variable inequality constraint (SVIC) and requires special treatment when a trajectory lies on the boundary of the state space* (which is a compact set) (see Chap. 6 of [21], [7], [14], and [20]). In order to be able to apply existing results without inessential complications we consider only two X-force target types.

McIntyre and Paiewonsky [20] remarked in 1967 that "the optimal control problem with state space constraints does not appear to be well understood." Moreover, this author [27] has recently pointed out that a necessary condition of optimality involving the multiplier associated with a first order SVIC (see [14]) which has been given by Gamkrelidze (see Chap. 6 in [21]) has apparently been overlooked in the applications literature (engineering, operations research).** Gamkrelidze's multiplier condition will play an essential roll in characterizing an optimal allocation policy for the above problem with nonautonomous dynamics. This is the first application of the theory of SVIC's to such a tactical allocation problem in the Lanchester theory of combat to the best of the author's knowledge.

In this paper we consider only the development of necessary conditions of optimality that hold on an extremal battle trajectory for the above problem (50). The synthesis of an extremal trajectory from these conditions follows in a fashion similar to the developments of Sec. 3.2. The determination of domains of controllability (see [26]) for various terminal states of combat does not appear to be mathematically tractable except in special cases, and even then it is very messy. Needless to say, the determination of the optimal battle trajectory (and hence the corresponding optimal fire distribution policy) from among the extremal candidates is even more complicated and, hence, is beyond the scope of our present examination.

---

*An example of this is when $x_1 = 0$ for $t_e \leq t \leq t_l$.

**However, this multiplier condition has been adequately handled by mathematicians in many pieces of purely mathematical research [10], [11], [12], and [22].

4.1. Development of Basic Necessary Conditions of Optimality

The Hamiltonian for problem (50) is given by [7] and [14]

$$H(t, x_i, p_i, \phi) = p_1(t)\{-\phi a_1(t)y + r_1(t)\} + p_2(t)\{-(1-\phi)a_2(t)y + r_2(t)\} \tag{51}$$
$$+ p_3(t)\{-b_1(t)x_1 - b_2(t)x_2 + s(t)\} + \mu_1(t)\{-\phi a_1(t)y + r_1(t)\} + \mu_2(t)\{-(1-\phi)a_2(t)y + r_2(t)\},$$

where

$$\mu_i(t) \begin{cases} = 0 & \text{for } x_i > 0, \\ \geq 0 & \text{for } x_i = 0. \end{cases}$$

The adjoint system of differential equations for the dual variables is

$$\frac{dp_1}{dt} = b_1(t)p_3, \tag{52}$$

$$\frac{dp_2}{dt} = b_2(t)p_3, \tag{53}$$

$$\frac{dp_3}{dt} = \phi^* a_1(t)(p_1 + \mu_1) + (1-\phi^*)a_2(t)(p_2 + \mu_2). \tag{54}$$

At a corner interior to the state space the corner conditions (12) and (13) hold with $n=2$. When x_1, $x_2 > 0$, the control law is determined by the maximum principle, and one readily obtains results similar to those presented in [30].

On a constrained on which $x_1(t) = 0$ (and $x_2 > 0$) for $t_e \leq t \leq t_l$ the control is determined by $dx_1/dt = 0$ so that

$$\phi^*(t) = \frac{r_1(t)}{a_1(t)y} \qquad \text{for } t_e < t < t_l. \tag{55}$$

The multiplier $\mu_1(t)$ is determined by the condition that $\partial H/\partial\phi = 0$ and hence

$$\mu_1(t) = \frac{1}{a_1(t)}\{a_2(t)p_2(t) - a_1(t)p_1(t)\}. \tag{56}$$

The condition* that $\mu_1(t) \geq 0$ yields that on a constrained subarc we must have

$$a_1(t)(-p_1(t)) \geq a_2(t)(-p_2(t)). \tag{57}$$

Similar to previous developments, let us consider the *special case in which*

$$a_i(t) = k_{a_i}h(t) \qquad \text{for } i = 1, 2. \tag{58}$$

Then (56) becomes

$$\mu_1(t) = \frac{1}{k_{a_1}}\{k_{a_2}p_2(t) - k_{a_1}p_1(t)\}. \tag{59}$$

Differentiating (59) with respect to time and combining the result with (52), (53), and (58), we find that

---

*This corresponds to taking the dual variables to be continuous at an exit from a constrained subarc [14].

$$\mu_1(t) = \frac{p_3(t)}{a_1(t)} \{a_2(t)b_2(t) - a_1(t)b_1(t)\}, \tag{60}$$

so that Gamkrelidze's multiplier condition [21], [27]

$$\dot{\mu}_1(t) \leq 0, \tag{61}$$

is only satisfied on a constrained subarc with $x_1 = 0$ and $p_3(t) > 0$ when

$$a_1(t)b_1(t) \geq a_2(t)b_2(t). \tag{62}$$

Furthermore, there are corner conditions that must be satisfied upon entering to and exiting from a constrained subarc. Let t_e denote the time of entry to the constrained subarc and t_l denote the exit time. This situation is depicted in Figure 1. Furthermore, let t_e^- denote a left-hand limit, i.e.,

$$t_e^- = \lim_{\substack{t \to t_e \\ t \leq t_e}} t.$$

Applying the results* of McIntyre and Paiewonsky [20], we have the following conditions at an *entrance corner* to a constrained subarc on which $x_1 = 0$ for a finite interval of time $t_e \leq t \leq t_l$

$$p_1(t_e^-) = p_1(t_e^+) + \mu_1(t_e^+), \tag{63}$$

$$p_2(t_e^-) = p_2(t_e^+), \tag{64}$$

$$p_3(t_e^-) = p_3(t_e^+), \tag{65}$$

and

$$p_1(t_e^-) = \frac{a_2(t_e)}{a_1(t_e)} p_2(t_e^-). \tag{66}$$

At an *exit corner* to such a constrained subarc we have

$$p_i(t_l^-) = p_i(t_l^+) \qquad \text{for } i = 1, 2, 3, \tag{67}$$

and

$$H^*(t_l^-) = H^*(t_l^+), \tag{68}$$

so that

$$\mu_2(t_l^-) = \mu_2(t_l^+) = 0. \tag{69}$$

4.2. A Special Case of the Above Results: The Isbell and Marlow Fire Programming Problem

When

$$a_i(t) = a_i \qquad \text{and } b_i(t) = b_i \qquad \text{for } i = 1, 2,$$

$$r_1(t) = r_2(t) = s(t) \equiv 0, \tag{70}$$

and

$$T_1 = +\infty,$$

the above problem (50) reduces to the fire programming problem first studied by Isbell and Marlow

* Apparently there is a sign error in Equation (52) on p. 410 of [20].

[13]. (Additionally, one needs to consider explicitly various terminal states of battle (see [26]).) We did not employ the theory of SVIC's in our previous work [26], but (following the original work of Isbell and Marlow [13], which antedated the development of the theory of SVIC's) considered a "reduced" game in which $x_1, x_2 > 0$ for $t < t_e$. Moreover, this heuristic approach will not work for the problem at hand (50) due to the consideration of both replacements and also variable attrition-rate coefficients.

For the Isbell and Marlow fire programming problem in which (70) holds, it is seen from the above condition (62) that a necessary condition for it to be optimal to annihilate x_1 at $t_e < T$ (or, equivalently, to have $x_1(t) = 0$ for $t_e \leq t \leq T$ with $t_e < T$) is that (since $p_3(t) > 0$ for $t < T$)

$$a_1 b_1 \geq a_2 b_2 \tag{71}$$

Conversely, if one makes the nonrestrictive assumption that $a_1 b_1 > a_2 b_2$ then an optimal fire distribution policy does not result in $x_2(t) = 0$ for $t < T$. In the future we will discuss further how to use the theory of SVIC's to develop more extensive results on when it is not optimal to drive a force level to zero, the optimal order in which to annihilate target types, etc.

5. COMMENTS

In this paper we have attempted to show how some seemingly complex optimal control problems may readily yield relatively simple analytic solutions when special mathematical structures of the problem are exploited. We first pointed out such possibilities in a recent note [25] (in which we inadvertently rediscovered some results apparently first observed by B. O. Koopman [28]). Although the solution to matrix differential equations (such as those encountered in Sec. 3 and 4) is, in general, messy at best when specific solutions are required, the occurrence of a "bang-bang" optimal control in these optimization problems reduced the complexity of the solution appreciably.

Additionally, we have shown that the theory of state variable inequality constraints (SVIC's) (especially Gamkrelidze's multiplier condition) is essential for the treatment of the nonnegativity of force levels in fire distribution problems in the Lanchester theory of combat. The author is aware of no previous work relating the theory of SVIC's to such allocation problems in the Lanchester theory of combat. (More remarkably, the author in performing this research discovered that Gambrelidze's multiplier condition had been omitted in a significant portion of the applied literature [27].) For the problem with variable attrition-rate coefficients and replacements considered in Sec. 4, we saw that simple analytic results regarding when it is not optimal to drive a force level to zero were possible when we assumed that $a_i(t) = k_{a_i} h(t)$. In this case, a necessary condition for it to be optimal to have $x_1(t) = 0$ for $t_e \leq t \leq t_l$ was that $a_1(t) b_1(t) \geq a_2(t) b_2(t)$. Moreover, the developments of Sec. 4.1 must be modified when there are more than two X-force target types, and we shall discuss this in the future.

6. SUMMARY

In this paper we have considered two extensions of some problems discussed in a companion paper [30]. We have shown how optimal controls may be synthesized in a simple fashion when special mathematical structures of the problems at hand are exploited. Some of the results developed here justify the results given without proof in [30] for a two-on-one combat problem with variable attrition-rate coefficients. Additionally, we have shown how to apply the theory of state variable inequality constraints (SVIC's) to fire distribution problems in the Lanchester theory of combat in order to treat the nonnegativity of force levels. As a result of our investigations, we have hopefully laid the groundwork for studying more general and complex structures in the optimization of combat dynamics.

REFERENCES

[1] C. Barfoot, "The Lanchester Attrition-Rate Coefficient: Some Comments on Seth Bonder's Paper and a Suggested Alternative Method," Operations Research *17*, 888–894 (1969).

[2] V. Boltyanskii, *Mathematical Methods of Optimal Control* (Holt Rinehart and Winston, New York, 1971).

[3] S. Bonder, "Theory for Weapons System Analysis," Proceedings U.S. Army Operations Research Symposium (1965), pp. 111–128.

[4] S. Bonder, "The Lanchester Attrition-Rate Coefficient," Operations Research, *15*, 221–232 (1967).

[5] S. Bonder, "The Mean Lanchester Attrition Rate," Operations Research, *18*, 179–181 (1970).

[6] H. Brackney, "The Dynamics of Military Combat,"Operations Research *7*, 30–44 (1959).

[7] A. Bryson and Y. C. Ho, *Applied Optimal Control* (Blaisdell Publishing Company, Waltham, Massachusetts, 1969).

[8] C. von Clausewitz, *On war*, edited with an introduction by A. Rapoport (Penguin Books Ltd., Harmondsworth, Middlesex, England, 1968).

[9] L. Fan, *The Continuous Maximum Principle* (John Wiley & Sons, Inc., New York, 1966).

[10] J. Funk and E. Gilbert, "Some Sufficient Conditions for Optimality in Control Problems with State Space Constraints," SIAM J. Control *8*, 498–504 (1970).

[11] T. Guinn, "The Problem of Bounded Space Coordinates as a Problem of Hestenes," SIAM J. Control *3*, 181–190 (1965).

[12] M. Hestenes, *Calculus of Variations and Optimal Control Theory* (John Wiley & Sons, Inc., New York, 1966).

[13] J. Isbell and W. Marlow, "Methods of Mathematical Tactics," Logistics Papers, *14*, The George Washington University Logistics Research Project (Sept. 1956).

[14] D. Jacobson, M. Lele, and J. Speyer, "New Necessary Conditions of Optimality for Control Problems with State-Variable Inequality Constraints," J. Math. Anal. Appl. *35*, 255–284 (1971).

[15] C. Johnson, "Singular Solutions in Problems of Optimal Control," pp. 209–267 in *Advances in Control Systems*, C. Leondes (Ed.) (Academic Press, New York, 1965), vol. 2.

[16] C. Johnson and J. Gibson, "Singular Solutions in Problems of Optimal Control," IEEE Trans. on Automatic Control, *AC–8*, 4–15 (1963).

[17] S. Kimbleton, "Attrition Rates for Weapons with Markov-Dependent Fire," Operations Research *19*, 698–706 (1971).

[18] E. Lee and L. Markus, *Foundations of Optimal Control Theory* (John Wiley & Sons, Inc., New York, 1967).

[19] B. H. Liddell Hart, *Strategy* (Frederick A. Praeger, Inc., New York, 1967) (Second Revised Edition).

[20] J. McIntyre and B. Paiewonsky, "On Optimal Control with Bounded State Variables," pp. 389–419 in *Advances in Control Systems*, C. Leondes (Ed.) (Academic Press, New York, 1967), Vol. 5, pp. 389–419.

[21] L. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mishchenko, *The Mathematical Theory of Optimal Processes* (Interscience, New York, 1962).

[22] I. B. Russak, "On Problems with Bounded State Variables," J. Opt. Th. App. *5*, 114–157 (1970).

[23] J. Rustagi and R. Laitinen, "Moment Estimation in a Markov-Dependent Firing Distribution," Operations Research *18*, 918–923 (1970).

[24] J. Rustagi and R. Srivastava, "Parameter Estimation in a Markov-Dependent Firing Distribution," Operations Research *16*, 1222–1227 (1968).

[25] J. Taylor, "A Note on the Solution to Lanchester-Type Equations with Variable Coefficients," Operations Research *19*, 709–712 (1971).

[26] J. Taylor, "On the Isbell and Marlow Fire Programming Problem," Nav. Res. Log. Quart. *19*, 539–556 (1972).

[27] J. Taylor, "Comments on a Multiplier Condition for Problems with State Variable Inequality Constraints," IEEE Trans. on Automatic Control, Vol. *AC–17*, 743–744 (1972).

[28] J. Taylor, "Comments on 'A Note on the Solution to Lanchester-Type Equations with Variable Coefficients,'" Operations Research *20*, 1194–1195 (1972).

[29] J. Taylor, "Solving Lanchester-Type Equations for 'Modern Warfare' with Variable Coefficients," Operations Research, *22*, 756–770 (1974).

[30] J. Taylor, "Lanchester-Type Models of Warfare and Optimal Control," Nav. Res. Log. Quart. *21*, 79–106 (1974).

[31] J. Taylor, "Target Selection in Lanchester Combat: Linear-Law Attrition Process," Nav. Res. Log. Quart. *20*, 673–697 (1973).

[32] H. Weiss, "Lanchester-Type Models of Warfare," in *Proceedings First International Conference Operational Research* (John Wiley & Sons, Inc., New York, 1957), pp. 82–99.

UPPER AND LOWER BOUNDS FOR A PATTERN BOMBING PROBLEM

Alan R. Washburn
Naval Postgraduate School
Monterey, California

ABSTRACT

This paper includes two simple analytic formulas for kill probability that are applicable in circumstances where shots should be fired in a pattern. The two formulas bracket the maximum kill probability achievable with an optimal pattern. The upper bound corresponds to an optimal nonfeasible pattern, and the lower bound to a nonoptimal feasible pattern.

INTRODUCTION

The general subject of "coverage problems" has received considerable effort in the literature. Reference [5] contains a good bibliography. This paper is devoted to a somewhat narrower topic, namely, "What sort of a pattern should be used when several shots have to be fired simultaneously at a given target?" This topic also has a long history. The question has been answered rigorously only when the number of simultaneous shots is small, and the methods used do not permit easy extension to larger numbers. The approach here will be to develop upper and lower bounds on the kill probability (expected fraction killed), keeping in mind the primary goal of developing an analytical model that is applicable when the number of shots is large. Two sample applications are included at the end of this paper.

HISTORICAL BACKGROUND

World War II gave impetus to work in the field of pattern bombing, with the applications being to artillery, antiaircraft fire, and air support. Svesnikov [12] developed an approximation to the highest possible kill probability obtainable when firing N cookie-cutter shots with normal aiming errors at a point target in one dimension. His approximation is actually an upper bound, and it is his approach that will be used in obtaining an upper bound here. Much later, Marsaglia [8] found the exact solution for $N=2$ by graphical methods.

Von Neumann [10] found an exact analytical solution for $N=2$ by "diffusing" the kill relationship between the target and the bomb; specifically, he assumed that the probability that a bomb landing r units away will kill the target is of the form $\exp(-Kr^2)$, rather than the usual cookie-cutter function. Diffuse bombs generally produce "cleaner" analytical results than cookie-cutter bombs. Either assumption can be defended as being closer to reality, depending on the application. Dobbie [3] has proposed a sequence of functions that is diffuse on one end and cookie cutter in the limit. The diffuse assumption will be made here.

Our lower bound will be obtained by examining a particular aiming strategy; namely, "aim all shots at the same point, but degrade the delivery accuracy so that the shots actually fall in some sort

of random 'pattern'." Since any degradation of delivery accuracy will result in a lower bound, the problem is to find that degradation that results in the greatest lower bound. Circumstances where the optimal amount of degradation is 0 were investigated by Kolmogorov [12]. The subject is also taken up by Merritt and King [9], who give a method for approximately determining the optimal round-to-round dispersion in circumstances very similar to those treated here.

Many other authors have investigated the problem of optimal aiming [1], [4], [6], [7], [13], [15]. In most cases, the methods used result in approximate expressions for kill probability, with it sometimes being the case that the approximation is neither obviously high nor obviously low, and often being the case that no bound on the error is available. In this paper, the kill probability will be bracketed between upper and lower bounds.

ASSUMPTIONS

1) There is a circular normal (σ_E) error common to all shots.

2) There is a single area target with value density

$$V(x, y) = (V_0/2\pi\sigma_T^2) \exp(-(x^2+y^2)/2\sigma_T^2) \qquad \text{for all } (x, y).$$

3) The position of the center of the target is known to within a circular normal (σ_L) error.

4) The round-to-round dispersion (independent error for each shot) is circular normal (σ_I).

5) Any element of the target will be killed by a shot r units from the element with probability $R \exp(-r^2/2c^2)$, where $R \leq 1$, and all shots are independent for each element (no cumulative damage).

ANALYSIS

The assumptions are stated as though the target were an area target, and this will be a convenient frame of reference to maintain; however, $\sigma_T = 0$ corresponds to a point target. Since it has been shown elsewhere [6] that σ_E, σ_T, and σ_L are important only through the sum $\sigma_E^2+\sigma_T^2+\sigma_L^2$, it is no restriction to assume $\sigma_E=\sigma_L=0$. This result is intuitively reasonable because uncertainty of location should have the effect of "spreading the target out," and whether an error common to all shots is attributed to the marksman or the cartographer is unimportant—the sum is all that matters. Finally, since the problem will be discussed in terms of the fraction of the target killed (or not killed), the total target value V_0 will not appear below.

Let (x_i, y_i) be the point at which the ith shot is aimed. This shot will kill the point (x, y) with probability

$$(1) \qquad P_i(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\frac{1}{2\pi\sigma_I^2} \exp\left(-\frac{(u-x_i)^2+(v-y_i)^2}{2\sigma_I^2}\right)\right] \left[R \exp\left(-\frac{(x-u)^2+(y-u)^2}{2c^2}\right)\right] du dv,$$

where (u, v) is the impact point of the ith round. This integral is easily shown [15] to be

$$(2) \qquad P_i(x, y) = R\frac{c^2}{c^2+\sigma_I^2} \exp\left(-\frac{(x-x_i)^2+(y-y_i)^2}{2(\sigma_I^2+c^2)}\right).$$

The simple form of (2) is the primary reason for assumption 5.

The probability that the point (x, y) will not be killed by any shot is $\prod_{i=1}^{N}(1-P_i(x, y))$, and it follows that the *average fraction* of the target value not killed (miss probability in the case of a point target) is

$$P_M=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\frac{1}{2\pi\sigma_T^2}\exp\left(-\frac{x^2+y^2}{2\sigma_T^2}\right)\right]\left[\prod_{i=1}^{N}(1-P_i(x, y))\right]dxdy. \tag{3}$$

By substituting $p=R\,\frac{c^2}{c^2+\sigma_I^2}$, $u=x/\sigma_T$, $v=y/\sigma_T$, $S^2=\sigma_T^2/(\sigma_I^2+c^2)$, $u_i=x_i/\sigma_T$, and $v_i=y_i/\sigma_T$, (3) becomes

$$P_M=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}(u^2+v^2)\right)\prod_{i=1}^{N}\left(1-p\exp\left[-\frac{S^2}{2}\{(u-u_i)^2+(v-v_i)^2\}\right]\right)dudv. \tag{4}$$

THE UPPER BOUND

We follow here the method of Svesnikov. Equation (4) can be written

$$P_M=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left[-\left\{\frac{1}{2}(u^2+v^2)+D(u, v)\right\}\right]dudv, \text{ where} \tag{5}$$

$$D(u, v)=-\sum_{i=1}^{N}\ln\left(1-p\exp\left[-\frac{S^2}{2}\{(u-u_i)^2+(v-v_i)^2\}\right]\right). \tag{6}$$

Whatever the u_i and v_i are, $D(u, v)$ is a nonnegative, continuous function. Further,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}D(u, v)\,dudv=-\sum_{i=1}^{N}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ln\left(1-p\exp\left[-\frac{S^2}{2}\{(u-u_i)^2+(v-v_i)^2\}\right]\right)dudv. \tag{7}$$

Each of the N integrals in the right hand side of (7) is independent of u_i and v_i, on account of the infinite limits. Therefore, all of the u_i and v_i can be set to 0 for purposes of integration. If polar coordinates are then introduced, the result is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}D(u, v)\,dudv=2\pi N\int_0^{\infty}-\ln\left(1-p\exp\left(-S^2r^2/2\right)\right)rdr \tag{8}$$

$$=(2\pi N/S^2)\int_0^{\infty}-\ln\left(1-p\exp\left(-v\right)\right)dv. \tag{9}$$

Let

$$a\equiv RNc^2/\sigma_T^2. \tag{10}$$

The quantity a can be interpreted as the ratio of "expected area covered with widely spaced shots" to "target area."

Also, let

$$I(p) \equiv p^{-1} \int_0^{\infty} -\ln(1-p\exp(-v))dv = p^{-1} \int_0^{p} -u^{-1}\ln(1-u)du, \tag{11}$$

where the second equality in (11) is obtained by substituting $u=p\exp(-v)$.

We can now rewrite (9) as

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} D(u,v)dudv = (2\pi Np/S^2)I(p) = 2\pi aI(p). \tag{12}$$

Let $\hat{S}$ be the set of all functions $D(u,v)$ that are nonnegative, continuous, and have volume $2\pi aI(p)$, so that $\hat{S}$ includes all functions of the form (6). It follows that

$$P_M \geq \inf_{D(u,v)\in\hat{S}} \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-\left\{\frac{1}{2}(u^2+v^2)+D(u,v)\right\}\right]dudv. \tag{13}$$

Svesnikov [12] used a variational argument to conclude that the right hand side of (13) is

$$(1+\sqrt{2aI(p)})\exp(-\sqrt{2aI(p)}) \equiv 1-\bar{P}_K, \tag{14}$$

where $\bar{P}_K$ is the upper bound on the kill probability. This is the "square root formula" that has since been obtained by other authors in other ways [6].

The function $\bar{P}_K$ depends on the round-to-round dispersion σ_I only through the function $I(p)$, which means that the dependence is not very strong. In fact,

$$1=I(0)\leq I(p)\leq I(1)=\pi^2/6\approx 1.645 \tag{15}$$

so that, if $p_K=1-p_M$,

$$p_K \leq 1-(1+\sqrt{2a\pi^2/6})\exp(-\sqrt{2a\pi^2/6}) \tag{16}$$

in all cases. The function $I(p)$ is shown in Figure 1. Figure 2 shows $\bar{P}_K$ as a function of a for $p=0$ and 1, being formula (16) in the last case. Figure 2 also shows a lower bound that will now be derived.

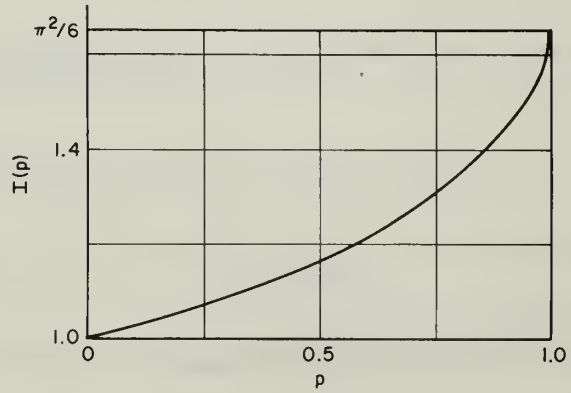


FIGURE 1. The function $I(p)$.

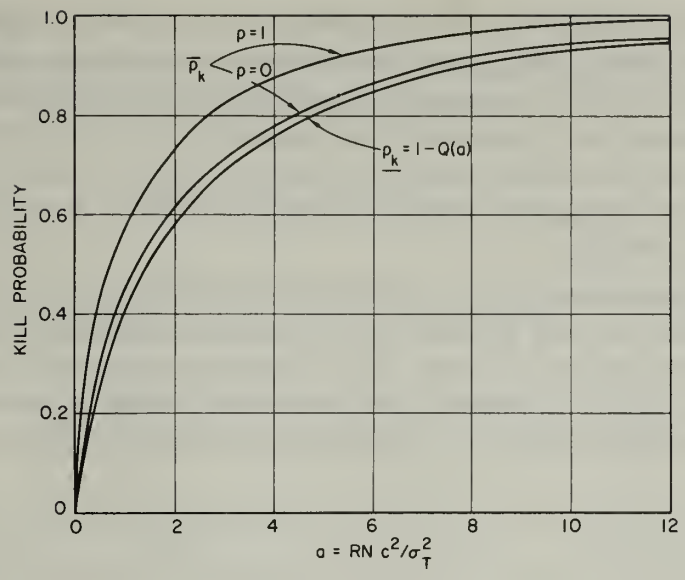


FIGURE 2. Upper and lower bounds on the kill probability.

THE LOWER BOUND

The prospect of actually optimizing (4) with the u_i and v_i is not encouraging. An attempt to equate derivatives to 0 would almost certainly lead to nothing but $2N$ very messy equations, except at $u_1 = \ldots = u_N = v_1 = \ldots = v_N = 0$, where all the derivatives are 0. Kolmogorov [12] was able to establish that this "pattern" is at least locally optimal for certain parameter values in a different problem by investigating second order conditions. Such a result would be of little use here, however, since the primary application is supposed to be to situations where the weapons should obviously be spread out in some manner or other.

A second approach would be to guess a pattern to within a scale factor, and then search for a scale factor that provides minimum p_M within the class of patterns. For example, when $N=3$, the shots should "obviously" be aimed in a triangle, and the only real issue is the size of the triangle. This approach is feasible for small N, but the type of pattern to use is not obvious when N is large. Even when $N=4$, it is not clear whether "square" is superior to "triangle-plus-one-in-the-middle." Of course, any feasible pattern provides a lower bound of some sort, but it is nonetheless evident that this approach would become unwieldy for large N.

A third approach (the one that will be used here) is to select the aim points as independent random variables from some common distribution. This approach holds analytical promise because round-to-round dispersion basically serves that function anyway, and is easy to handle. Picking the aim points at random is equivalent to changing the dispersion; in fact, if (U_i, V_i) is the random aim point of the ith shot, and if (X_i, Y_i) is the dispersion, then the only important distribution is the common distribution of the N independent "equivalent dispersions" (X_i+U_i, Y_i+V_i). There are now two choices for subsequent analysis,

1) find the optimal distribution of (U_i, V_i), perhaps within some class of distributions,

2) optimize something more aggregated than the distribution of (U_i, V_i), such as the distribution of (X_i+U_i, Y_i+V_i), or even the convolution of this function with the damage function.

The reason for considering the second approach is that it permits the application of variational techniques that often lead to simple, closed form solutions for the kill probability [13], [15]. Unfortunately, the optimized distribution or convolution would not be physically realizable in this problem, being 0 everywhere except on a bounded set. It is impossible to add two independent random variables, one of which (the dispersion) is normal, to obtain a sum that is bounded with probability one. This means that the second approach will not result in a lower bound on the kill probability, since it may involve infeasible strategies.

Approach 1) is simpler in concept than approach 2), but harder to carry out unless the class of distributions is suitably restricted. We shall restrict the class of distributions for (U_i, V_i) to circular normal distributions centered on the origin. This will simplify computations, while still resulting in a lower bound on the optimal kill probability, because the only changes in formula (4) are that all the u_i and v_i become 0, and the parameter σ_I is replaced by $\sigma \equiv$ standard deviation of total dispersion. The result is

$$P_M \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-(u^2+v^2)/2\right)\left(1-p_1 \exp\left(-S_1^2(u^2+v^2)/2\right)\right)^N du\,dv, \tag{17}$$

where $p_1 = R\,c^2/(c^2+\sigma^2)$, $S_1^2 = \sigma_T^2/(c^2+\sigma^2)$, and $\sigma \geq \sigma_I$. Changing to polar coordinates,

$$P_M \leq \int_0^{\infty} \exp\left(-r^2/2\right)\left(1-p_1 \exp\left(-S_1^2 r^2/2\right)\right)^N r\,dr. \tag{18}$$

Let $b = 1/S_1^2$, and substitute $u = p_1 \exp\left(-S_1^2 r^2/2\right)$. Then (18) becomes

$$P_M \leq b p_1^{-b} \int_0^{p_1} u^{b-1}(1-u)^N du. \tag{19}$$

The integral is an incomplete beta-function, as reported by Breaux [2]. One further approximation is in order. Since $1-u \leq \exp(-u)$, (19) can be replaced by

$$P_M \leq b p_1^{-b} \int_0^{P_1} u^{b-1} \exp\left(-Nu\right) du, \tag{20}$$

and the substitution $v = Nu$, $a = Np_1 b$ gives

$$P_M \leq b(a/b)^{-b} \int_0^{a/b} v^{b-1} \exp\left(-v\right) dv \equiv Q(b, a). \tag{21}$$

The integral is now an incomplete Gamma function [9]. The quantity $a \equiv RNc^2/\sigma_T^2$ does not depend on σ. When $\sigma = \sigma_I$, $b = (c^2+\sigma_I^2)/\sigma_T^2 \equiv b_{\min}$, so

$$P_M \leq \inf_{b \geq b_{\min}} Q(b, a). \tag{22}$$

THE FUNCTION $Q(b, a)$

Since $e^{-v} \leq 1$, it follows from the definition of $Q(b, a)$ in (21) that $0 \leq Q(b, a) \leq 1$. Also, a power series expansion for e^{-v} followed by term integration reveals that

$$Q(b,a) = \sum_{j=0}^{\infty} \frac{(-a/b)^j}{j!} \frac{b}{b+j}, \tag{23}$$

which is well suited for computation when a/b is not large. Also, (23) shows that $Q(b, 0) = 1$.

An integration of (22) by parts shows that

$$Q(b,a) = (b/a)^b \int_0^{a=b} u^b \exp(-u)du + \exp(-a/b). \tag{24}$$

Equation (24) makes it clear that $\lim_{b \to 0} Q(b, a) = 1$. Also, since $\exp(-a/b) \leq \exp(-u) \leq 1$,

$$\exp(-a/b)(1+a/(b(b+1))) \leq Q(b,a) \leq \exp(-a/b)+a/(b(b+1)), \tag{25}$$

and $\lim_{b \to \infty} Q(b, a) = 1$. Since $Q(b, a)$ is not identically one for $a > 0$, it must have at least one local minimum in b over the interval $0 < b < \infty$.

Computation has shown that a unique minimum of $Q(b, a)$ occurs at some $b = b^*(a)$ for $a > 0$. The function $\underline{P}_K(a) = Q(b^*(a), a)$ is shown as the lower bound on the kill probability in Figure 2. This "lower bound" is obviously valid only if $b^*(a) \geq b_{\min}$; otherwise, the lower bound should be $Q(b_{\min}, a)$. The function $Q(b, a)$ has been tabulated by Pearson [11], although anyone with access to a small computer would probably find it easier to use (23). See also the report [14].

The function $a/b^*(a) \equiv N(a)$ is shown in Figure 3. Since $a/b_{\min} = RNc^2/(c^2+\sigma_I^2) = Np$, the lower

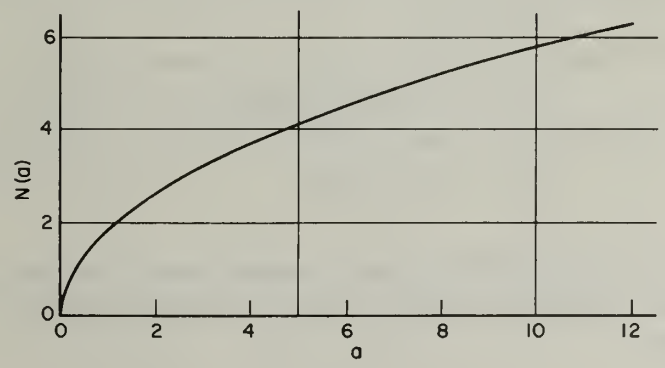


FIGURE 3. Lower bound valid if $Np \geq N(a)$.

bound shown in Figure 2 is valid as long as Np is larger than $N(a)$. If R is interpreted as an in-flight reliability, or a defense penetration probability, or some combination of both, then NR is the expected number of shots arriving at the target, and Np is a slightly smaller number. In other words, the lower bound in Figure 2 will not be valid for attacks with a few large weapons.

EXAMPLES

Example 1:

$$N=4, \quad c=0.5, \quad \sigma_T=1, \quad \sigma_I=0, \quad R=0.7.$$

In this case, it is possible to compute the actual optimum kill probability, assuming that the shots should be placed either in a "square" or "triangle-plus-one-in-the-middle" pattern, by numerically integrating formula (4) and doing a one-dimensional search. The square with unit side turns out to be superior, and $P_K=0.380$. Since $a=p=0.7$ in this example, Figure 1 shows $I(0.7)=1.27$, and formula (14) gives $\bar{P}_K=0.389$. Since $Np=2.8>N(a)=1.61$, the lower bound in Figure 2 is valid, and $\underline{P}_K=0.323$. The upper bound is evidently closer to the truth than the lower bound in this example. The two bounds can be brought closer together by increasing σ_I until $Np=N(a)$, which occurs when $\sigma_I=0.43$ if the other parameters are as given (the optimizing σ is 0.43 regardless of σ_I, but $\sigma=0.43$ is feasible only if $\sigma_I \leq 0.43$). The lower bound would then remain at 0.323, but the shrinking of p to 0.402 would cause the upper bound to shrink to 0.357. This will always be the case; the two bounds will be closest together when the lower bound is just on the verge of being invalid.

Example 2:

$$N=40, \quad c=0.5, \quad \sigma_T=1, \quad \sigma_I=0.2.$$

Also, each weapon has a nonreprogrammable reliability of 0.7, and a probability 1/7 of penetrating the (well defended) target's defenses. If all of the penetration events are assumed independent, then the 0.7 and 1/7 numbers can be multiplied to give $R=0.1$. Proceeding as usual,

$$a=(0.1)(40)(0.5)^2/(1)^2=1$$

$$p=0.1(0.5)^2/(0.5^2+0.2^2)=0.0862$$

$$I(p)=1.02$$

$$\bar{P}_K=0.417$$

$$Np=3.448>N(a)=1.912$$

$$P_K=1-Q(a)=0.402$$

An estimate of 0.41 for the kill probability with an optimal pattern could not be in error by more than 0.008. The computational difficulties involved in actually finding the optimal aim points for each of the 40 shots are obvious.

SUMMARY

The procedure developed here provides upper and lower bounds on the optimal kill probability with N shots. The bounds are often close together, particularly for large engagements. Application involves only one dimensional table look-ups, even though the number of input parameters can be quite large. This simplicity of results is made possible primarily by the assumptions of a normal target and a "diffused" bomb, which assumptions are important limitations to application.

REFERENCES

[1] Bressel, Charles N., "Expected Target Damage for Pattern Firing," Operations Research *19*, 655–667 (1971).

[2] Breaux, H. J., "A Note on Two Coverage Problems for Multiple Shots," Operations Research *16*, 1239–1242 (1968).

[3] "Probability-of-Damage Problems of Frequent Occurrence," Operations Evaluation Group Study No. 626 (1959).

[4] Duncan, R. L., "Hit Probabilities for Multiple Weapons Systems," Soc. Indus. Appl. Math, Rev. *6*, 111–114 (1964).

[5] Eckler, A. R., "A Survey of Coverage Problems Associated with Point and Area Targets," Technometrics *11*, 561–589 (1969).

[6] Galiano, R. J. and H. Everett, III, "Some Mathematical Relations for Probability of Kill," Lambda Paper 6 (1967), The Lambda Corporation.

[7] Grubbs, Frank E., "Expected Target Damage for a Salvo of Rounds with Elliptical Normal Delivery and Damage Functions," Operations Research *16*, 1021–1026 (1968).

[8] Marsaglia, G., "Some Problems Involving Circular and Spherical Targets," Operations Research *13*, 18–27 (1965).

[9] Merritt, H. L. and F. G. King, "Calculation of Engagement Kill Probability when the Aiming Errors of the Rounds Fired in the Engagement are not Independent," presented at the Second Anti-aircraft Fire Control Working Conference, Ballistic Research Laboratory Report 932 (1955).

[10] Von Neumann, J., Appendix to "Optimum Spacing of Bombs or Shots in the Presence of Systematic Errors," L. S. Dederick and R. H. Kent, Ballistic Research Laboratory Report No. 241 (3 July 1941).

[11] Pearson, K., *Tables of the Incomplete Gamma Function* (Cambridge University Press, 1934).

[12] Svesnikov, A. A., "Collection of Articles on the Theory of Firing," (A. N. Kolmogorov, ed.), Translation T–14, The RAND Corporation (1948).

[13] Walsh, J. E., "Optimum Ammunition Properties for Salvos," Operations Research *4*, 204–212 (1956).

[14] Washburn, A. R., "Approximations for a Pattern Bombing Problem," NPS55WS72051A, Naval Postgraduate School, Monterey, California (1972).

[15] Weiss, H. K., "Methods for Computing the Effectiveness of Area Weapons," Ballistic Research Laboratory Report 879 (1953).

THE MINIMUM SPHERE COVERING A CONVEX POLYHEDRON

Jack Elzinga
Johns Hopkins University
and
Donald Hearn
University of Florida

ABSTRACT

A finite algorithm is given for finding the smallest sphere enclosing a convex polyhedron in E^n described by a given system of linear equalities or inequalities. Extreme points of the polyhedron, and minimum spheres enclosing them, are generated in a systematic manner until the optimum is attained.

1. INTRODUCTION

In a recent paper in this journal Nair and Chandrasekaran [4] give a geometrical procedure for constructing the minimum circle containing a given convex polyhedron in E^2. Their procedure is designed for solving the problem by hand. They also give a primal formulation of the problem which is valid in E^n and useful provided the extreme points of the polyhedron are known.

The problem is

$$\min_{(s,\,x)} s \tag{1}$$

subject to $$s \geqq (y_i - x)^t (y_i - x) \qquad i = 1, 2, \ldots, m,$$

where y_i are extreme points of

$$S = \{y \mid Ay = b,\ y \geqq 0\}.$$

Note that $s^{1/2}$ and x are the radius and center of the sphere. Nair and Chandrasekaran show that (1) may be reformulated as a quadratic program by the change of variable $\lambda = s - x^t x$. The dual of (1) is also a quadratic program as we show in [2] and can be solved for the minimum covering sphere (MCS) by a decomposable procedure [2]; again provided the y_i are known. In the following section we give an algorithm for obtaining the MCS of S when the extreme points are not known. Specifically, the procedure generates extreme points and successively larger spheres covering them in a systematic manner which intuitively is more efficient than generating all of the extreme points (potentially a very large number). The problem is a fairly difficult one involving convex quadratic maximization, itself a problem not well solved. The procedure is efficient, up to a point, namely in constructing a sphere covering S, although not necessarily the minimal one. At this point the optimal sphere is known to lie between two known spheres.

First we give the algorithm and then a series of comments on the individual steps. We note that this procedure is probably too involved for applications in E^2, where all the extreme points may be obtained geometrically and then solving for the minimum circle geometrically [3]. Its potential usefulness lies in higher dimensional applications such as finding "centers" of convex polyhedrons (see [5, 7]).

2. THE ALGORITHM

We employ the following notation:

$$S=\{x|Ax=b,\ x\geqq 0\}$$

$$||\cdot||_p = l_p \text{ norm}$$

$$f(x, c)=||x-c||_2^2.$$

If x^k is an extreme point of S, we partition x^k as (basic variables, nonbasic variables)$=(x_B^k, x_N^k)$ and similarly partition $A=(B_k, N_k)$, where B_k is the associated basis matrix. Row vectors are denoted by prime ($'$).

The steps of the algorithm are:

(0) Determine one extreme point of S. Label it x^1 with associated basis B_1. Set $k=1$.

(1) Determine the MCS for $x^1, x^2, \ldots, x^k$ with center c^k and radius r_k.

(2) Let $L=\{k|f(x^k, c^k)=r_k^2\}$. Select $k\epsilon L$ and compute the row vector

$$\frac{\partial f'}{\partial x_N^k}=2[(x_N^k-c_N^k)'-(x_B^k-c_B^k)'B_k^{-1}N_k]$$

evaluated at $x^k=(x_B^k, x_N^k)=(x_B^k, 0)$. If some component of this vector is nonnegative then perform a Simplex pivot on the associated nonbasic variable to generate a new extreme point x^{k+1}. Set $k=k+1$ and go to 1. Otherwise, select another $k\epsilon L$ and repeat this step. Go to 3 if $\dfrac{\partial f}{\partial x_N^k}<0$ for all $k\epsilon L$.

(3) Solve $\max_x ||x-c^k||_1$ s.t. $x\epsilon S$ and denote its extreme point solution by $\bar{x}$ with basis $\bar{B}$. If $f(\bar{x}, c^k)>r_k^2$ let $x^{k+1}=\bar{x}, B_{k+1}=\bar{B}$, $k=k+1$ and go to 1. Otherwise go to 4. (See comment 5.)

(4) Solve $\max_x f(x, c^k)$ s.t. $x\epsilon S$ and denote its extreme point solution by $\hat{x}$ with basis $\hat{B}$. If $f(\hat{x}, c^k)\leqq r_k^2$, stop. Otherwise, let $x^{k+1}=\hat{x}$, $B_{k+1}=\hat{B}$, $k=k+1$ and go to 1.

3. COMMENTS

We note the following:

(1) In step 0 the initial extreme point is generated by the standard "phase 1" method of linear programming (or an extreme point may be known *a priori*).

(2) The MCS of step 1 is obtained by quadratic programming and is decomposable as explained in Reference [2].

(3) The set L identifies those known extreme points which lie on the current sphere (as opposed to within it). If from any such point there is an edge of S emanating outward from the current sphere this will be identified in step 2 and the neighboring external extreme point along this edge will be generated by the pivot. This step is analagous to "pricing out" in linear programming, because the components of $\dfrac{\partial f'}{\partial x_N^k}$ are the directional derivatives of f along the edges from x^k.

(4) Note that if some extreme point is degenerate all basic solutions representing it may be generated, but the algorithm will not cycle, as may occur in linear programming, since the only effect is duplication of vectors. These are recognized in step 2 by $x^{k+1}=x^k$, in which case step 2 is continued rather than going to 1.

(5) The steps 0–3 are relatively efficient (step 3 is a linear program), but step 4 requires some enumerative procedure such as branch-and-bound. Step 3 is motivated by the fact that rectilinear distance between two points is an upper bound on Euclidean distance. We note that when step 3 terminates with $f(\bar{x}, c^k) \leqq r_k^2$ then the sphere of radius $|\bar{x}-c^k|_1$ with center at c^k is a covering sphere for S although not necessarily the minimal one. This follows from the fact that if some extreme point of S is exterior to this larger sphere then it would be further (in the rectilinear sense) from c^k then $\bar{x}$. The optimal sphere lies between these two spheres.

(6) The maximization of a convex quadratic function required in step 4 has been addressed by Ritter [6] and Cabot and Francis [1] who propose a branch-and-bound procedure. The latter technique is the natural one to apply in step 4 because the function $g(x)=||x-c^k||_1^2$ is an upper bound for $f(x, c^k)$ for any $x \epsilon S$ and r_k^2 is a lower bound.

(7) If the set S is expressed by

$$S=\{x|\bar{A}x \leqq b,\ x \geqq 0\} \subset E^n,$$

where $\bar{A}$ is $(m \times n)$ then adding slack vectors raises the dimension of the space from E^n to E^{n+m} since the tableaux will be $m \times (m+n)$. Application of the procedure would generate a sphere in E^{n+m}. Rather then construct the desired sphere in E^n from this "larger" sphere we modify the procedure as follows. Using the $m \times (n+m)$ tableaux, all of the steps are as before except the calculation of directional derivatives in step 2. We use the same notation, except with the understanding that each vector is in E^n so that $f(x, c^k): E^n \rightarrow E^1$. For some x^k let d_j be an edge of S emanating from x^k. The derivative of f along this edge is $\nabla_x f(x^k, c^k)d_j=2(x^k-c^k)d_j$ with d_j being obtained from the nonbasic columns of the associated tableau. Specifically, for nonbasic column $\bar{A}_j$, d_j is of the following form (subject to possible rearrangement of components):

$$d_j=\begin{bmatrix} B^{-1}\bar{A}_j \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \epsilon E^n$$

with the 1 in position j.

REFERENCES

[1] Cabot, A. V. and R. L. Francis "Solving Certain Nonconvex Quadratic Minimization Problems by Ranking Extreme Points," Operation Research, *18*, 82–86 (1970).

[2] Elzinga, J. and D. W. Hearn, "The Minimum Covering Sphere Problem," Management Science *19*, 96–104 (1972).

[3] Elzinga, J. and D. W. Hearn, "Geometrical Solutions for Some Minimax Location Problems," Transportation Science, Vol. 6, No. 4 (1972).

[4] Nair, K. P. K. and R. Chandrasekaran, "Optimal Location of a Single Service Center of Certain Types," Nav. Res. Log. Quart., *18*, 503–510 (1971).

[5] Nemhauser, G. L. and W. B. Widhelm, "A Modified Linear Program for Columnar Methods in Mathematical Programming," Operations Research *19*, 1051–1060 (1971).

[6] Ritter, K., "Stationary Points of Quadratic Maximum Problems," Z. Wahrscheinlichkeitstheorie Verw., Geb *4*, 149–158 (1965).

[7] Wilde, D. J., *Optimum Seeking Methods*, Prentice-Hall, Englewood Cliffs, N.J., 1964.

AN ITERATIVE PROCEDURE FOR NONDISCOUNTED DISCRETE-TIME MARKOV DECISIONS

Gary J. Koehler
Northwestern University
Andrew B. Whinston
Purdue University
Gordon P. Wright
Purdue University

ABSTRACT

This paper considers the problem of computing, by iterative methods, optimal policies for Markov decision processes. The policies computed are optimal for all sufficiently small interest rates.

INTRODUCTION

This paper utilizes the linear programming procedure given by Manne [3] for finding gain optimal policies and the procedure given by Denardo [1] for finding bias optimal policies which, by repeated application, gives policies optimal for all sufficiently small interest rates. The problem to be addressed is the efficient solution of the corresponding linear programs.

Consider the following linear program and its dual:

Primal:	Max $c'x$	Dual:	Min $b'\pi$
	subject to:		subject to:
	$Ax = b$		$A'\pi \geqq c$
	$x \geqq 0$		$b \geqq 0$
	$b \geqq 0$		

where A' is the transpose of A, A is m by n, b and π are m by 1, and x and c are m by 1. Let J be a set of column identifications of A and A_J a submatrix of A consisting of the columns of A listed in J.

Consider a primal basis J. Let

$$A'_J = R_J - Q_J,$$

such that R_J is nonsingular and

(1) $$x \geqq y \text{ implies that } R_J^{-1}Q_J x \geqq R_J^{-1}Q_J y$$

(2) $$p(R_J^{-1}Q_J) < 1,$$

where $p(\cdot)$ is the spectral radius. If there exists an R and Q for each primal feasible basis J then iterative methods may be used to solve the dual linear program [2, 5].

An iterative procedure can be specified by using superscripts on vectors to denote iteration counts and superscripts on matrices to denote a power. Let π^0 be an arbitrary vector, J an initial basis, $n=0$, and parameters K and δ be specified. The iterative procedure is:

STEP 1:

$$\pi^{n+K+1} = (R_J^{-1}Q_J)^{K+1}\,\pi^n + \sum_{l=0}^{K} (R_J^{-1}Q_J)^l R_J^{-1}c_J.$$

STEP 2: Choose another basis, M, such that

$$R_M^{-1}Q_M\pi^{n+K+1} + R_M^{-1}c_M = \operatorname*{Max}_J R_J^{-1}Q_J\,\pi^{n+K+1} + R_J^{-1}c_J.$$

STEP 3: If $$||R_M^{-1}Q_M\pi^{n+K+1} + R_M^{-1}c_M - R_J^{-1}Q_J\pi^{n+K+1} - R_J^{-1}c_J|| \leqq \delta.$$

stop; otherwise set n to $n+K+1$,

$$\pi^n \text{ to } R_M^{-1}Q_M\pi^{n+K+1} + R_M^{-1}c_M$$

and label M by J. Go to Step 1.

δ is a solution tolerance and K is the number of refinements in Step 1. Actually, a much more comprehensive iterative procedure can be employed. The reader is encouraged to peruse Koehler, et al., [2] for more details.

The advantage of using iterative methods in linear programming stem primarily from deletion of the need for storing and maintaining a basis inverse. Storage requirements are reduced, computational effort is reduced, and arithmetic precision is easily controlled.

The concept of using iterative methods in linear programming is now applied to the problem of finding a gain optimal policy, a bias optimal policy, or a policy optimal for all sufficiently small interest rates in a discrete time Markov decision with an infinite horizon. Since such problems may be large and, at the very least, may require the repetitive solution of linear programs, the use of iterative methods for solving linear programs appears desirable.

NONDISCOUNTED DISCRETE TIME MARKOV DECISIONS

At each stage a system is observed to be in one of a finite set of states S. For any state $i \epsilon S$ observed at time $t(=0, 1, 2, \ldots)$, a decision $a \epsilon A_i$ is made. Each set A_i is finite. The outcome of the decision is an immediate expected reward $c_i(a)$ and a probability of moving to state j in the next stage. The conditional probability is denoted by $P_{ij}(a)$. A stationary nonrandomized policy δ is a vector valued func-

tion that specifies a decision for each state. That is, for each $i \epsilon S$, $\delta(i) \epsilon A_i$. Let F be the set of all nonrandomized stationary policies. Then each policy $\delta \epsilon F$ has associated with it an $N \times 1$ vector of expected rewards $c(\delta)$ and an $N \times N$ matrix of transition probabilities $P(\delta)$ where N is the number of states in S. Throughout this paper only problems possessing irreducible state spaces for every $\delta \epsilon F$ are considered.

For a discounted infinite horizon problem, the vector, $\pi(\delta)$, represents the present values for each state and is given by:

$$\pi(\delta) = \sum_{t=0}^{\infty} \alpha^t P(\delta)^t c(\delta),$$

where α is the discount factor and $0 < \alpha < 1$.

A nondiscounted infinite horizon problem may be solved by considering the limiting process of $\pi(\delta)$ as $\alpha \to 1^-$. Miller and Veinott [4] have given a complete characterization of 1^- optimal policies and a policy iterative method for finding such policies.

When only policies optimizing the gain or average expected reward are desired, Manne's [3] linear programming formulation may be used. For finding bias optimal policies or, by repeated application, finding policies optimal for all sufficiently small interest rates, Denardo's [1] linear programming procedures may be used.

In this paper, the linear programming methods given by Denardo [1] and Manne [3] are modified to allow efficient iterative methods [2] in their solution.

GAIN OPTIMALITY

When only policies optimizing the gain or average expected reward are desired, Manne's [3] linear programming formulation may be used. The formulation is:

$$\text{Max} \sum_{i \epsilon S} \sum_{a \epsilon A_i} c_i(a) x_{ia}$$

Subject to

(1) $$\sum_{i \epsilon S} \sum_{a \epsilon A_i} x_{ia} = 1.$$

(2) $$\sum_{i \epsilon S} \sum_{a \epsilon A_i} B_{ia} x_{ia} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$x_{ia} \geqq 0 \qquad i \epsilon S, \qquad a \epsilon A_i,$$

where

(3) $$B_{ia} = \begin{pmatrix} -P_{i1}(a) \\ 1 - P_{ii}(a) \\ \vdots \\ -P_{i,N-1}(a) \end{pmatrix}$$

Note that (2) is pre-Leontief [6], as the vector B_{ia} has one positive element when $i < N$. Now consider a 1 by $N-1$ vector y, such that

$$y'B_{ia} \geqq 1 \qquad \text{For all } i < N \text{ and } a \epsilon A_i. \tag{4}$$

Note from (3) that (4) is equivalent to

$$y_i \geqq 1 + \sum_{j=1}^{N-1} P_{ij}(a) y_i. \tag{5}$$

So a vector satisfying (4) has y_i as the longest expected time until state N is observed given state i. Premultiply (2) by y and subtracting from (1) gives

$$\sum_{i \epsilon S} \sum_{a \epsilon A_i} (1 - y'B_{ia}) x_{ia} = 1. \tag{6}$$

The coefficient of x_{ia} in (6) is, from (4), nonpositive when $i < N$. Also, since $B_{Na} < 0$ and $y_i \geqq 1$, the coefficient of x_{Na} in (6) is strictly positive. So together (2) and (6) have exactly one positive coefficient associated with each variable x_{ia}. This equivalent system is pre-Leontief. Since any feasible pre-Leontief system is equivalent to a Leontief system [6] and any Leontief system (after suitable permutations) has feasible basis which can be split to yield asymptotic monotone contractions [2], the equivalent formulation given by (2) and (6) can be solved by iterative techniques. Furthermore, the determination of the vector y involves solving a pre-Leontief substitution problem. That is, minimize

$$\sum_{i=1}^{N-1} y_i$$

subject to (5). Note though that the solution tolerance must be quite small so that no positive entries occur in the coefficients of x_{ia} for $i \neq N$ of (6).

BIAS OPTIMAL POLICIES

Denardo [1] has demonstrated that a bias optimal policy may be found by restricting the policy space and altering the cost structure of the gain optimal policies. After such changes, another linear program is solved to determine the bias optimal policies.

Consider the linear program given by (2) and (6). In compact form, the problem (7) may be written as

$$\text{Max } c'x, \tag{7}$$

subject to

$$Bx = b$$

$$x \geqq 0.$$

Let

$$N = \underset{i \epsilon S}{X} \{(a, i) : a \epsilon A_i\}.$$

Also, let the set K be a subset of N constructed using the method of Denardo's Problem II [1]. Furthermore, let $\bar{c}$ be the vector of altered costs according to Denardo's Problem II. Then Equation (7) may be modified to

$$\text{Max } \bar{c}'x,$$

subject to

$$Bx = b$$

$$x \geqq 0;$$

and only basis B_J, $J \epsilon K \subseteq N$ are considered. Since the constraint set has not been changed, the problem is still a pre-Leontief structured linear program and may be solved using iterative methods [2]. The restricted basis entry is the only imposition on the algorithm (which, incidentally, simplifies the computational procedure). Hence, a bias optimal policy may be found by changing the cost structure and basis entry rules while leaving the constraint set unchanged. Furthermore, iterative methods may be used for the determination of the bias optimal policy.

OPTIMAL POLICIES FOR SMALL INTEREST RATES

As Denardo [1] has pointed out, by repeated application of the above procedure, the optimal policy for all sufficiently small interest rates can be determined.

It is important to note that the constraint set remains pre-Leontief once Manne's original problem has been converted to a pre-Leontief structured linear program. In all cases iterative procedures for solving linear programs may be employed.

ACKNOWLEDGMENT

The authors thank Eric V. Denardo for suggesting significant improvements in the interpretation and presentation of the original version of this paper.

REFERENCES

[1] Denardo, Eric V., "Computing a Bias-Optimal Policy in a Discrete-Time Markov Decision Problem," Operations Research *18*, 279–289 (1970).

[2] Koehler, Gary J., Andrew B. Whinston, and Gordon P. Wright, "The Solution of Large Scale Leontief Substitution Systems," submitted for publication.

[3] Manne, Alan S., "Linear Programming and Sequential Decisions," Management Science *6*, 259–267 (1960).

[4] Miller, Bruce L. and Arthur F. Veinott, Jr., "Discrete Dynamic Programming with a Small Interest Rate," Ann. Math. Stat. *40*, 366–370 (1969).

[5] Veinott, Arthur F., Jr., "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria," Ann. Math. Stat. *40*, 1635–1660 (1969).

[6] Veinott, Arthur F., Jr., "Extreme Points of Leontief Substitution Systems," Linear Algebra and Its Applications *1*, 181–194 (1968).

A NOTE ON MATHEMATICAL ASPECTS OF THE $3 \times n$ JOB-SHOP SEQUENCING PROBLEM

Wlodzimierz Szwarc

School of Business Administration
University of Wisconsin-Milwaukee
Milwaukee, Wisconsin

One can offer the following simplification of the solution method for the case $B_i \geq A_j$ for all i and j (see [1], p. 147).

STEP 1: Solve the $2 \times n$ BC problem. Let $p = (p_1, \ldots, p_n)$ be the optimal schedule.

STEP 2: Find $\omega \overset{df}{=} \{i \mid A_{p_i} < A_{p_1}\}$.

$$\text{Define } p^{(i)} = (p_i, p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n).$$

STEP 3: Among the schedules p and $p^{(i)}$, $i \epsilon \omega$ find the sequence with the minimum processing time.

This is the optimal solution of the problem.

EXAMPLE 1

	A	B	C
1	6	8	4
2	8	9	9
3	5	7	12
4	7	10	5
5	4	11	8
6	3	8	10

We solve the two machine BC problem where $p = (3, 6, 2, 5, 4, 1)$ is the optimal schedule. Find $\omega = \{i \mid A_{p_i} < 5 = A_3\}$. Hence $\omega = \{5, 6\}$. According to step 3 we consider three sequences 362541. 536241, and 632541.

	A	B	C
3	5	7^{12}	12^{24}
6	3^{8}	8^{20}	10^{34}
2	8^{16}	9^{29}	9^{43}
5	4^{20}	11^{40}	8^{51}
4	7^{27}	10^{50}	5^{56}
1	6^{33}	8^{58}	4^{62}

	A	B	C
5	4	11^{15}	8^{23}
3	5^{9}	7^{22}	12^{35}
6	3^{12}	8^{30}	10^{45}
2	8^{20}	9^{39}	9^{54}
4	7^{27}	10^{49}	5^{59}
1	6^{33}	8^{57}	4^{63}

	A	B	C
6	3	8^{11}	10^{21}
3	5^{8}	7^{18}	12^{33}
2	8^{16}	9^{27}	9^{42}
5	4^{20}	11^{38}	8^{50}
4	7^{27}	10^{48}	5^{55}
1	6^{33}	8^{56}	4^{60}

Hence sequence 632541 with the smallest processing time ($=60$) is optimal.
The solution method for case 2b ($B_i \geqslant C_j$ for all i, j) is symmetrical.

* * *

Errata: Formulas (6) and (7) of [1] should read

(6) $$g(p) = \ldots = A_{p_1} + \max_{1 \leqslant v \leqslant n} H_v$$

(7) $$A_{p_1} + \min_{\vec{p}} \max_{1 \leqslant v \leqslant n} H_v$$

REFERENCES

[1] Szwarc W., "Mathematical Aspects of the $3 \times n$ Job-Shop Sequencing Problem," Nav. Res. Log. Quart. **21**, 145–153 (1974).

INDEX TO VOLUME 21

☆U.S. GOVERNMENT PRINTING OFFICE: 1975-581-588/2

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

DECEMBER 1974
VOL. 21, NO. 4
NAVSO P-1278

CONTENTS

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217